NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

AUTOMATIC KEYFRAME SUMMARIZATION OF USER-GENERATED VIDEO

by

Eric C. Eckstrand

June 2014

Thesis Co-Advisors: Mathias Kolsch
Ronald Giachetti

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704–0188

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	06-20-2014	Master's Thesis 10-20-2013 to 06-20-2014

4. TITLE AND SUBTITLE

AUTOMATIC KEYFRAME SUMMARIZATION OF USER-GENERATED VIDEO

5. FUNDING NUMBERS

6. AUTHOR(S)

Eric C. Eckstrand

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Postgraduate School
Monterey, CA 93943

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

The views expressed in this document are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol Number: N/A.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(maximum 200 words)*

The explosive growth of user-generated video presents opportunities and challenges. The videos may possess valuable information that was once unavailable. On the other hand, the information may be buried or difficult to access with traditional methods. Automatic keyframe video summarization technologies exist that attempt to address this problem. A keyframe summary can often be viewed quicker than the underlying video. However, a theoretical framework for objectively assessing keyframe summary quality has been absent. The work presented here bridges this gap by presenting a semantically high-level, stakeholder-centered evaluation framework. The framework can capture common stakeholder concerns and quantitatively measure the extent to which they are satisfied by keyframe summaries. With this framework, keyframe summary stakeholders and algorithm developers can now identify when success has been achieved. This work also develops a number of novel keyframe summarization algorithms and shows, using the evaluation framework, that they outperform baseline methods.

14. SUBJECT TERM

Video Summarization, Video Abstraction, Keyframes, Keyframe Summary Evaluation, User-Generated Video, Video Segmentation, Keyframe Selection

15. NUMBER OF PAGES 91

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UU

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

AUTOMATIC KEYFRAME SUMMARIZATION OF USER-GENERATED VIDEO

Eric C. Eckstrand
Lieutenant, United States Navy
B.S., United States Naval Academy, 2004
M.S., University of Delaware, 2006

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN SYSTEMS ENGINEERING

from the

NAVAL POSTGRADUATE SCHOOL
June 2014

Author: Eric C. Eckstrand

Approved by: Mathias Kolsch
Thesis Co-Advisor

Ronald Giachetti
Thesis Co-Advisor

Clifford Whitcomb
Chair, Department of Systems Engineering

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The explosive growth of user-generated video presents opportunities and challenges. The videos may possess valuable information that was once unavailable. On the other hand, the information may be buried or difficult to access with traditional methods. Automatic keyframe video summarization technologies exist that attempt to address this problem. A keyframe summary can often be viewed quicker than the underlying video. However, a theoretical framework for objectively assessing keyframe summary quality has been absent. The work presented here bridges this gap by presenting a semantically high-level, stakeholder-centered evaluation framework. The framework can capture common stakeholder concerns and quantitatively measure the extent to which they are satisfied by keyframe summaries. With this framework, keyframe summary stakeholders and algorithm developers can now identify when success has been achieved. This work also develops a number of novel keyframe summarization algorithms and shows, using the evaluation framework, that they outperform baseline methods.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

NPS Naval Postgraduate School
USG United States Government
ZB zettabytes = 10^7 bytes
IDC International Data Corporation
RGB red-green-blue
PCA pricipal components analysis
GMM Gaussian mixture model
MRF Markov random field
SVM support vector machine
SRD shot reconstruction degree
AOC artifact of concern
POC property of concern
CC combinatorial concern
TC temporal concern
ROC receiver operating characteristic
AUC area under the curve
LIBSVM Library for Support Vector Machines
SBD shot boundary detection
VATIC Video Annotation Tool Irvine California

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgements

I would like thank my advisors, Mathias Kolsch and Ronald Giachetti, for their patience, encouragement, and guidance throughout the development of this research work. I would also like to acknowledge the indispensable assistance provided by Tom Batcha and Kyoung Min Lee. Finally, I would like to thank my parents, Philip and Cindy, for their overall support and encouragement.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Introduction

The amount of data created annually is growing at an exponential rate. The International Data Corporation (IDC) estimated that it grew by 48 percent from 2011 to 2012 to 2.7 zettabytes (ZB) [1]. They project that data will continue to grow at a rate of about 40 percent per year to 2020 [2].

Videos certainly contribute to a portion of the annual growth rate. For example, "100 hours of video are uploaded to YouTube every minute" [3]. This presents opportunities and challenges. The videos may possess valuable information that was once unavailable. On the other hand, the information may be buried or difficult to access with traditional methods. For example, "it would take an individual over 5 million years to watch the amount of video that will cross global IP networks each month in 2017. Every second, nearly a million minutes of video content will cross the network in 2017" [4].

The focus of this report is user-generated video, which is defined as video that is not professionally edited and, generally, has low production value. CNN News and Hollywood movies, for example, are not considered user-generated video. User-generated video may be produced using hand-held or body-worn camera devices, such as cell phones or Google Glass. User-generated video is also frequently published to social media websites, available for public consumption or for a small circle of friends.

1.1 Problem

As user-generated video grows, it is possible that it may be harnessed to achieve a number of desirable outcomes. For example, from a military perspective, the growth of camera sensors may expand visual coverage of the battlefield. However, situational awareness may not expand in proportion to camera sensor coverage. Decision-makers may become overwhelmed by the inundation of information.

Video may exist on social media websites that may be employed as evidence in a criminal investigation. Consider a group of police officers assigned with conducting a review of such

video. The size of the video collection may be large, and, with the available resources, it may be impractical, costly, or time consuming to conduct a thorough review.

The problem is not unique to military and law enforcement organizations. YouTube users, for example, watch videos for entertainment purposes. They may encounter a daunting amount of candidate videos in the process. A simple YouTube query of "cat," for example, returns over 32 million results [3]. Consumers have an interest in finding the most interesting content in the least amount of time. Content providers have an interest in satisfying these desires.

The problem of having limited resources to consume all of the desired video affects a large number and variety of people.

1.2 Approach

A number of approaches exist that address the problem. For example, additional personnel may be harnessed, or personnel skill may be enhanced, so that more video can be consumed in a given period of time. The focus of this report, however, is on a particular technological solution: keyframe video summarization.

If a video is viewed as a sequence of individual frames, then keyframe video summarization entails selecting a subset of frames from the entire video frame sequence. In this paper, the keyframes are automatically selected by a computer program. The resulting frames are then be viewed by a user.

1.2.1 Evaluation

Automatic keyframe summarization methods may be evaluated in a number of ways. For example, user studies may be conducted to determine the enjoyability or informativeness of a keyframe summary [5]. Keyframe summarization methods may also be evaluated objectively. For example, candidate keyframe summaries may be generated, and then compared to objective, ground truth summaries.

There are advantages and disadvantages associated with either evaluation approach discussed above. User studies are labor intensive. Objective measures, on the other hand, are computed automatically. However, objective measures may only be an approximation of

human judgment [5].

Keyframe summarization methods are evaluated with objective measures in this research, and an evaluation by user studies is deferred to future work.

1.3 Key Contributions

The contributions of this research are twofold. First, this work presents a semantically high-level, stakeholder-centered evaluation framework. The framework can capture common stakeholder concerns and quantitatively measure the extent to which they are satisfied by keyframe summaries. Second, this work develops a number of novel keyframe summarization algorithms and shows, using the evaluation framework, that they outperform baseline methods.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 2:
Related Work

The field of video summarization can be characterized along a number of dimensions: summary display format, and the approaches used for video segmentation, keyframe selection, and summary evaluation [5]. The author discusses related research along these dimensions below.

2.1 Keyframe Summaries Versus Video Skims

Generally speaking, two main video summarization display formats exist: keyframe summaries and video skims. In the keyframe format a video is represented as a set of frames, or as a storyboard [6–15]. Each frame of the storyboard is a still image. Sound and motion from the original video are absent. A video skim, on the other hand, is a set of video clips extracted from the target video [16–20]. For example, a movie trailer may be considered a video skim [5].

There are trade-offs associated with both formats, including size and expressiveness. Keyframe summaries are generally more compact, that is, they can represent the same visual concepts with fewer frames. Video skims, on the other hand, may possess non-visual information, such as human speech and sound [5].

The keyframe summarization format was chosen for this research because summary compactness is a priority. Video skims may be considered in future work. Consequently, related keyframe summarization techniques are primarily discussed. However, some video skims are constructed from keyframes. In these cases, the related keyframe selection techniques will also be discussed.

Other, application-specific summary display formats exist. Rav-Acha, Pritch, and Peleg [21] consider surveillance video from a stationary camera. They compress the video by showing actions that occurred in the original video at different points in time, simultaneously. They also avoid action collisions, that is, two actions cannot occur in the same physical space at the same time in the summary. The author's research is not constrained

to such a specific type of video, so techniques associated with this display format are not discussed.

2.2 Segmentation Technologies

Segmentation approaches are often employed in video summarization. Here, a video is first temporally segmented into meaningful units, or sequences of consecutive frames. Next, keyframes are extracted from each unit. The type, location, number, and size of the discovered units themselves provide valuable summarization information. Additionally, the frames within a particular unit may be related, in terms low level characteristics, like color, or higher-level characteristics, like theme. The units' characteristics may then be exploited to improve the keyframe selection process [9, 13, 22–24].

The summarization approaches developed in this paper employ a number of segmentation strategies and the related ones are discussed below.

2.2.1 Shot Boundary Detection

Video is often segmented into its shots. A shot boundary is a change from one camera perspective to another. Of note, the change in camera perspective does not have to be immediate. For example, video may dissolve, fade, pan, or zoom from one shot to another. In this case, the boundary is not a hard shot boundary, but rather a soft shot boundary. Kobla, DeMenthon, and Doermann [25] propose a method to identify these less decisive shot boundaries, which is based on the red-green-blue (RGB) coefficients of the DC frame images. Other approaches exist, and Yuan et al. [26] cover them in an extensive shot boundary detection survey.

Shot boundary detection technology is employed in various ways to generate video summaries. Shots often represent a fundamental semantic unit of a video sequence [27]. If one thinks of a video's scenes as chapters in a book, then its shots may be considered paragraphs, and its frames may be considered the sentences. Some simple summarization methods identify a video's shots first and then generate a keyframe summary by selecting the first, middle, and/or last frame of each shot. Other methods select keyframes based on each shot's visual dynamics [5]. For example, Yeung and Liu [9] select the first frame of a shot. Subsequent frames within this shot are examined until a frame is encountered that sig-

nificantly differs from the first frame, or until the end of the shot is encountered. In the case where a significantly different frame is encountered, it is selected and subsequent frames are compared to this frame. This process continues until the end of the shot is reached. The difference between two frames is a function of luminance projections. Ferman and Tekalp [13], on the other hand, cluster the frames of each shot with fuzzy c-means clustering on color-based frame features. The frames closest to the cluster centers are selected as keyframes.

The summarization approaches developed in this paper do not explicitly employ shot boundary technology, but they do employ keyframe selection strategies similar to those presented above.

2.2.2 Motion Segmentation

In some domains video is unedited, or edited in a non-professional manner. It may be continuous in nature and possess no clear shot boundaries. Scene transitions may be smooth, occurring over longer periods of space and time. Additionally, the storyline may be less crafted or coherent when compared to professional cinema. As such, shot boundary technology may be of limited use. Lu and Grauman [22] introduce temporal segmentation technology that is tailored to video such as this.

Lu and Grauman [22] temporally segment a given video by first classifying each video frame as belonging to one of three motion-characteristic classes: static, in transit, and moving the head. Three one-vs.-rest support vector machine (SVM) classifiers are constructed, based on optical flow features, to perform this labeling [28]. Next, the video labeling is smoothed using a Markov random field (MRF). Consecutive frames with the same label are then designated as a distinct unit. The video is then organized into higher-level units called major events, based on the color similarity of the previously computed motion-characteristic units and their mutual influence. Finally, a keyframe selection algorithm is applied to each major event unit based on the labels of its motion-characteristic units and the objects contained within them [22].

The author develops a number of the summarization algorithms that employ a similar approach to Lu and Grauman [22]. The author's algorithms first classify each video frame. The number and type of classes vary, depending on the particular application. Next, the

algorithms smooth the classifications and organize consecutive frames of the same class into distinct units. Finally, keyframe selection is performed on the different units.

Other motion-based temporal segmentation algorithms exist. For example, Laganière et al. [29] compute the activity level of each frame. The activity level of a frame is a function of the number of significant Hessian spatio-temporal features present. Here, a feature is significant if the determinant of its Hessian matrix exceeds a predefined threshold. The video's activity level is then plotted over the video's frame sequence. Local maxima are identified, and consecutive frames that surround a local maximum frame, and have an activity level that exceeds a threshold value, are then designated as active. In this manner, a video is segmented in to active and inactive units. A video skim summary is then constructed by concatenating all of the active units.

The author develops summarization approaches that are similar to Laganière et al. [29]. Laganière et al. compute activity levels of each frame. The approaches in this work also compute activity levels of each frame, but the activity level is defined in a different manner. Here, the activity level is defined as the rate that the outdoor, indoor, man-made, natural, and greenery scenery is changing. Also like Laganière et al. the approaches presented in this work identify the locations of the local maxima in activity and perform keyframe selection based on them.

2.3 Keyframe Selection

Once a video is segmented into its units, perhaps by one of the segmentation technologies discussed above, keyframes must still be selected. The method may be as simple as selecting the first, middle, and/or last keyframe of each segment. Another basic approach is to draw a uniform sample of frames. Although simple, these approaches produce summaries which contain redundant and uninformative keyframes [5]. Simple approaches such as these serve as baselines for comparison against more sophisticated keyframe summarization techniques employed in this work.

Truong and Venkatesh [5] organize keyframe selection technologies based on general characteristics of their underlying selection algorithm, that is, whether they employ sufficient content change, equal temporal variance, maximum frame coverage, clustering, curve simplification, minimum correlation, sequence reconstruction error, or interesting-events meth-

ods. The clustering and interesting-events methods are most related to the author's research. These technologies, and other related keyframe selection technologies that do not fit nicely into the above characterization, are discussed below.

2.3.1 Clustering

In the clustering approach, a video frame is represented as a point in some feature space. Color histograms are often employed in this respect. The points are then clustered according to some clustering algorithm. For example, Girgensohn and Boreczky [30] employ hierarchical agglomerative clustering on frame color histograms. Yu et al. [31] also compute frame color histograms but employ kernel principal components analysis (PCA) to reduce the dimensionality of the feature space. They cluster the points with fuzzy c-means clustering. Gibson and Thomas [32] also reduce feature space dimensionality with PCA, but then cluster the points with a Gaussian Mixture Model (GMM) trained using Expectation Maximization.

Once the frames have been clustered, some approaches filter the clusters based on various criteria. For example, Zhuang et al. [33] remove clusters that do not exceed the average cluster size.

Finally, when clusters have been selected, various methods exist to select the keyframes from the clusters. A number of methods simply choose the point (frame) closest to the cluster center [31, 32]. Lee et al. [34], on the other hand, select the frame that has the highest importance score. Girgensohn and Boreczky [30] select a cluster's representative frame based on temporal constraints.

The algorithms presented in this paper employ k-means clustering on color histogram features, as well as features returned by various high-level object detectors. The algorithms also select frames based on how close their features are to the cluster centers.

2.3.2 Frame Aspects

A video frame may be characterized along a number of dimensions. For example, it may be characterized by its level of brightness or blur. At a higher semantic level, a frame may be characterized by its level of interestingness or representativeness. A frame may also be characterized by whether or not it possesses a certain object, such as a face or pedestrian.

The aspects that are of concern, however, depend on the stakeholder's needs. A number of keyframe selection techniques are centered around frame aspects. Existing approaches that are related to the approaches developed in this paper are discussed below.

Interestingness

Dufaux [35] defines frame interestingness as a measure of its motion activity, spatial activity, and the likelihood that people are present, as determined by skin-color and face detection technology. Each frame is scored based on this definition. The highest scoring frame is selected as the keyframe.

Representativeness

Kang and Hua [36] attempt to identify video frames that possesses a high degree of representativeness. However, unlike Dufaux [35], they avoid such a strict frame attribute definition. Instead, the authors observe that, given a particular video sequence, users assigned with manually selecting the most representative frames, tend to pick similar ones. The authors hypothesize that frame representativeness is highly related to image quality, user attention measure, and visual details. They attempt to learn this relationship by collecting manually-labeled examples of representative frames. From these training instances, a GMM is learned, which can automatically predict a new frame's representativeness based upon its underlying image quality, attention measure, and visual details. Keyframes are then selected based upon their computed level of representativeness.

Lee et al. [34] employ an approach similar to Kang and Hua [36] in that they learn a predictor for an ill-defined measure (representativeness) from manually-labeled training examples. However, in this case, the predictor's input is not a frame, but rather a frame region, or object, and importance, not representativeness, is being measured. More specifically, within a training frame set, they manually identify the important spatial regions within each frame. They hypothesize that a given region's importance can be predicted by a number of high and low-level features, such as region size, width, height, distance to the closest hand in the image, distance to the frame center, and likelihood that the region is a face. As such, they segment the training frames into a number of candidate regions, extract their features, and assign each candidate region an importance score equal to the ratio of the area of intersection to the area of union with the ground-truth important region. A linear regression model is then constructed from the training data, which is subsequently

used to predict the importance of new regions. Given a new video, the candidate regions of each frame are computed and then clustered. Keyframes are selected based on the most important region within each cluster.

Saliency

Jiang et al. [37] as in Lee et al. [34] and Kang and Hua [36], learn a predictor for an ill-defined measure from manually-labeled training examples. Like Lee et al. the predictor's input is a frame region, or object, but saliency is measured instead of importance. Saliency is defined as the degree to which a frame region, or object, draws the attention of the observer. Given a frame, they create a scaled image pyramid. Each image is then spatially segmented into its regions from which features are extracted. In contrast to Lee et al's work, semantically low-level features are employed that are based on regional contrast, region color and texture distributions, region size and position, and regional backgroundness. Based on these features, a previously-trained regression model returns a significance score. The regions, along with their scores, are then combined into a final saliency map, where more important pixels are assigned a greater grayscale intensity value than less important pixels.

Wang et al. [38] also employ lower-level features, such as multi-scale contrast, color spatial distribution, center-surround histogram, and region-based contrast to predict object saliency characteristics. In particular, they determine via a classification model whether or not a given image possesses a salient object. If it does, a single bounding box that encapsulates the salient object, is computed with a regression model. They restrict their approach to images with, at most, a single salient object.

Neither Jiang et al. [37] nor Wang et al. [38] investigate their technology's application to video summarization, in general, or keyframe selection, in particular. A number of approaches presented in this paper employ Jiang et al's salient object detection technology in a video summarization context. The approaches are then evaluated using the framework presented in this paper.

Humans

Humans are often the center of attention in videos, whether it be their presence, location, identity, actions, or relationships to other humans. In this regard, automatic human capture

technologies may be applied to the video summarization task. Moeslund and Granum [39] characterize human capture technologies along the dimensions of tracking, pose estimation, and recognition, among others. Tracking technology is particularly related to the work presented in this paper because it can help prepare videos for keyframe selection. Related face and pedestrian capture technologies are discussed below.

Faces

Given a picture of a query face, Sivic, Everingham, and Zisserman [40] retrieve pictures of the same person in a target video. For example, they use Julia Roberts as a query face and "Pretty Woman" as a target video. Particular instances of Julia Roberts' face that occur throughout the movie are returned. They accomplish this by first identifying all of the different faces within each frame using a frontal face detector. Next, the video is segmented into its shots, and, within each shot, faces of the same person are grouped by employing affine covariant region tracking combined with a single-link agglomerative grouping strategy. Faces of the same person are then grouped across shots using a similar strategy. Now, for a given person, for each instance of their face within the video (face-track), SIFT descriptors are computed for the eyes, nose, mouth, and mid-point between the eyes. An entire face-track is then represented as a distribution of these descriptors. Finally, given a query face, the face-track with the smallest chi-squared distance to the query face is selected.

Pedestrians

Andriluka et al. [41] attempt to detect and track multiple pedestrians in complex video environments. Initially, they detect a pedestrian's limb position and articulation with a parts-based object detection model. Based upon prior knowledge of body part articulations and their temporal consistency within a walking cycle, they construct a kinematic limb model. This model is then used to track the detected pedestrians across the video.

Sivic et al. [40] and Andriluka et al. [41] do not investigate their technology's application to video summarization. A number of approaches presented in this paper employ face and pedestrian detection technologies in a video summarization context. The approaches are then evaluated using the framework presented in this paper.

Scenery

Technology has been developed that classifies the genre of a video. Here, video genres are types of videos that shares similarities in content and structure. Many genres of video footage exist. Some examples include news, sports, movies, cartoons, and commercials. Rasheed et al. [42] classify video genres (comedy, action, drama, and horror) with low-level video statistics, such as average shot length, color variance, motion content, and lighting.

Li, Su, Xing, and Fei-Fei [43] present scene classification technology based on object banks. They create 12 scaled versions of a given image. 200 object detectors are then used to produce response maps at each scale. A three-level spatial image pyramid is created for each response map for a total of 36 scale-levels. For each object and for each scale-level, the response map is used to construct a feature histogram. The x-axis units correspond to one of the 200 objects. The y-axis units represent the cumulative response over all scale-levels for that object. They construct SVM scene classifiers based upon these histogram signatures, and achieve state-of-the-art results on UIUC-Sports [44] and MIT-Indoor [45] scene datasets.

Neither Rasheed et al. [42] nor Li et al. [43] investigate their technology's application to video summarization. A number of approaches presented in this paper employ Li et al's scene classification technology in a video summarization context. The approaches are then evaluated using the framework presented in this paper.

2.4 Evaluation Frameworks

Truong and Venkatesh [5] characterize evaluation methods based on whether they employ objective metrics, a results description, or user studies. Evaluations based on objective metrics are most related to the research presented here.

Many objective metrics have been developed to evaluate the quality of keyframe summaries. Chang et al. [46], for example, compute the semi-Hausdorff distance between a keyframe set and the original video segment from which it was drawn. The authors represent frames by their luminance projection vectors, but suggest color histograms may also be utilized. The authors submit that a keyframe summary with a small semi-Hausdorff distance has high fidelity.

Liu et al. [23], on the other hand, compute the shot reconstruction degree (SRD) of a shot's keyframes. The SRD indicates the degree to which the original shot can be reconstructed from the selected keyframes. SRD is computed, in part, by measuring the similarity between the frames in the keyframe set to all of the frames in the shot. The authors submit that if an original shot can be reconstructed from a set of keyframes well, then the keyframes capture the detailed motion dynamics of the shot.

Similarly, Lee and Kim [24] compute the amount of distortion among a shot's keyframes. To do so, they define a dissimilarity function that takes the features of consecutive keyframe pairs as input. Frame intensity and time are employed in this regard. However, the authors suggest that other dissimilarity measures based on frame texture, color, motion, or shape may also be employed. The authors compute the distortion of a keyframe summary by computing the area under the dissimilarity function for each consecutive keyframe pair and then summing the results. The authors suggest that keyframe summaries with low distortion are good temporal-visual representations of the underlying shot.

Liu and Kender [47] compute a keyframe summary's energy score by summing the visual distances between temporally adjacent frames in the summary. Keyframe summaries with high energy scores are considered well-distributed.

The objective metrics presented above are based on semantically low-level frame characteristics, such as intensity or luminance projection vectors. Consequently, they may fail to measure the degree to which a keyframe summary captures semantically high-level summarization concerns, such as objects or events. For example, the semi-Hausdorff distance, SRD, distortion, or energy of a keyframe summary may not be informative when a good keyframe summary is defined as one that contains "all of the football touchdowns." Quantitative measures of keyframe summary quality that capture semantically high-level concepts are needed.

Dufaux [35] attempts to generate keyframe summaries with respect to semantically high-level summarization concerns. A good keyframe summary, according to Dufaux, is one that is semantically meaningful and of high visual quality. However, Defaux does not elaborate on what is considered semantically high-level, nor does he attempt to quantify this definition with an objective measure. The evaluation framework that is proposed is

strictly subjective. However, Dufaux utilizes semantically high-level frame features in an attempt to create good keyframe summaries. In particular, Dufaux utilizes motion activity, spatial activity, skin-color, and face detection technologies to identify good keyframes. He compares the keyframe summaries generated with this method to those generated by selecting the middle frame from each of the shots.

Kim and Hwang [48] also attempt to generate keyframe summaries with respect to semantically high-level summarization concerns. Here, a good keyframe summary is one that captures an interesting object in its different poses or shapes. The authors limit the scope of their work to surveillance-style video, but once they determine the interesting object(s) for each video frame they compute its Hu moments. The first frame is selected as a keyframe. Successive keyframes are chosen sequentially based whether or not they are dissimilar enough to the previously selected keyframe. As in Dufaux [35], Kim and Hwang do not quantify the quality of the resulting keyframe summaries with objective metrics, but rather, subjectively describe their quality.

Lee et al. [34] also attempt to generate keyframe summaries with respect to semantically high-level summarization concerns and they quantify the quality of their keyframe summaries with objective metrics. The authors limit the scope of their work to ego-centric video, that is, the video camera is positioned on the head of the camera wearer. An important object is one that plays an important role in the context of the camera wearer's activities. A good keyframe summary is one that captures the important objects. The authors quantify this summarization concern by calculating the percentage of important objects captured by a keyframe summary. They compare their summarization method to a uniform sampling method and show that it is superior, that is, for any summary size, it finds a larger percentage of the important objects.

While Lee et al. [34] succeed in developing an objective metric that quantifies keyframe summary quality, with respect to a semantically high-level summarization concern, the summarization concern is still limited. Other valid concerns may exist, perhaps regarding the important object's size, geographic location, number, color, or some other property. The evaluation framework proposed by Lee et al. only considers important object presence.

The evaluation framework presented in this paper permits an evaluation of keyframe sum-

maries along a number of dimensions, including presence. In addition, the framework permits the evaluation of keyframe summaries with respect to a number of semantically high-level concepts, not just important people and objects.

CHAPTER 3:
Evaluation Framework

A good summary should capture the important aspects of its source content and ignore the superfluous aspects. Of course, what is important to one observer may be considered superfluous to another. Take, for example, a city building architect and a geologist. The architect may be interested in architectural elements, such as trusses, walls, facades, and foundations. The geologist, on the other hand, may be interested in strikes, dips, faults, and folds, among other geological structures. Given video of the Great Wall of China, an architect's summary may significantly differ from a geologist's. Consequently, the author proposes an evaluation framework that is stakeholder-centered, that is, the important aspects of a video are specified by the summary's stakeholders.

This chapter develops three tenets of a stakeholder-centered evaluation framework: the aspects that the stakeholders consider important, the method used to construct ground truth summaries based on these aspects, and the measures used to compare candidate keyframe summaries to their ground truth summaries.

3.1 Summarization Ideal

The collection of aspects that a stakeholder considers important is called the summarization ideal. The author does not attempt to identify everything that a stakeholder may find important, but rather, aspects related to a video's artifacts and their order of appearance. In sum, a stakeholder's summarization ideal consists of the artifacts of concern (AOCs), AOC properties of concern (POCs), AOC combinatorial concerns (CCs), and AOC temporal concern (TC). These concerns are discussed in detail below.

3.1.1 Artifacts of Concern

An artifact is "something characteristic of or resulting from a particular human institution, period, trend, or individual" [49]. Trusses, walls, facades, and foundations are examples of architectural artifacts. Pitches, catches, swings, and steals are baseball artifacts. Exposition, climax, and denouement are narrative artifacts. Artifacts that a stakeholder considers important are called AOCs.

3.1.2 Properties of Concern

AOCs may posses countless properties. For example, an AOC of car, may be characterized by its color, among other things. A person AOC has an identity property. Properties of an AOC that a stakeholder considers important are called POCs.

3.1.3 Combinatorial Concerns

AOCs may occur in different numbers or combinations within a video, which may hold special significance to a stakeholder. For example, given the AOCs of cat and dog, a portion of a video with three cats may be considered distinct from a portion with only one cat. Likewise, a portion of a video with both a cat and a dog may be considered distinct from a portion with one cat or just one dog. The AOC numbers or combinations that a stakeholder considers important are called the CCs.

3.1.4 Temporal Concern

An AOC may occur at different times, or frame locations, within a video. Despite being the same AOC, a stakeholder may consider each temporally unique AOC occurrence to be significant. If so, the stakeholder has a local TC. Otherwise, the stakeholder has a global TC.

Consider the case where the AOC is "Bob." Also, let us say that "Bob" occurs in segments of a video in the beginning, middle and end, which are separated by "non-Bob" portions. A stakeholder with a local TC would require "Bob" content from each segment, but a stakeholder with global TC would require "Bob" content from only one portion.

Of course, AOC temporal uniqueness may be defined in different ways, and the most appropriate measure may depend on the stakeholder's needs. Figure 3.1 depicts the summarization ideal.

The work presented here is limited to the above stakeholder concerns, but it is possible that a stakeholder may have concerns that are unrelated to artifacts or unrelated to the underlying content of the video. For example, a stakeholder may have limited computer storage or bandwidth capacity, and, therefore, may have a keyframe summary size concern. A stakeholder may desire video content at periodic intervals, regardless of the underlying

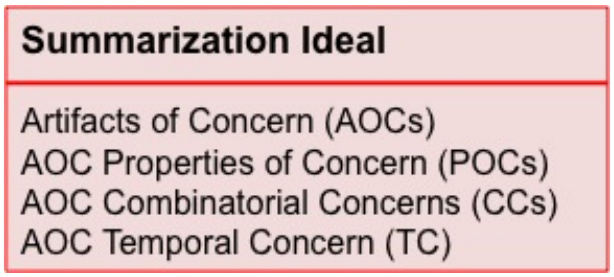

Figure 3.1: Summarization Ideal

content. The summarization ideal presented in this paper does not attempt to capture these type of concerns, but rather, content-related semantically high-level stakeholder concerns.

3.2 Summary Evaluation

Given a keyframe summary of a video, its quality can be determined by comparing it to the ground truth summary. Ground truth summary generation is discussed in this section. Then, metrics are introduced that can be used to compare keyframe summaries to their ground truth summaries.

3.2.1 Ground Truth Summary

The ground truth summary of a video can be derived by first labeling the video according to a given summarization ideal. Equivalent frames are grouped along the concerns specified in the summarization ideal. These groups comprise the ground truth summary. The video labeling and equivalent frame grouping processes are described below.

Video Labeling

First, for each frame, each AOC is labeled. At the very least, an AOC label consists of the AOC name and its respective frame number. An AOC label may also consist of POC values. For example, a triangle AOC may have an area POC value of 42 ft^2.

The AOC label data structure can be seen in Figure 3.2. Also, a sample of video labels can be seen in Figure 3.3, where the AOCs are pedestrians and cars, and no POCs are specified.

Equivalent Frames

Next, equivalent frames are grouped together. The collection of equivalent frame groups that result comprise the ground truth summary and it possesses all of the relevant video

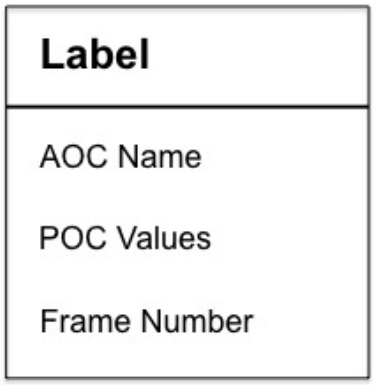

Figure 3.2: AOC Label Data Structure

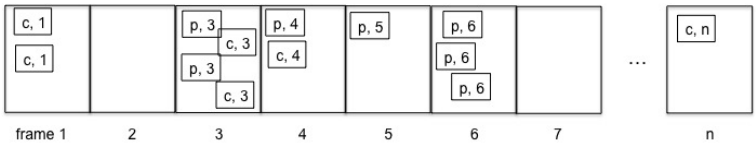

Figure 3.3: Video Labeling. AOCs: Pedestrian (p), Cars (c).

content, organized into distinct groups along the stakeholder's concerns.

In order group equivalent frames it is first necessary to define what it means for two frames to be frame equivalent. Frame equivalency depends on the summarization ideal, and example definitions based on various types summarization ideals are provided below.

Single-AOC, No POCs, No CCs, Global TC

Consider the summarization ideal where the AOC is pedestrian, no POCs or CCs are specified, and the TC is global. In this case, two frames are equivalent if they both possess a pedestrian, regardless of the pedestrian's identity. Figure 3.5 depicts the single equivalent frame group resulting from the video depicted in Figure 3.4. For a given summarization method, this means that if it picks at least one frame from the single equivalent frame group, then it has captured all of the desirable video content. However, if the summarization method picks frames from the single equivalent frame group, in excess of one frame, they are redundant.

Single-AOC, Single-POC, No CCs, Global TC

Now, consider the summarization ideal where the AOC is car, the POC is color, no CCs are specified, and the TC is global. In this case, two frames are equivalent if they possess a car

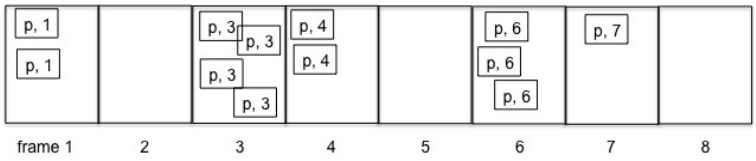

Figure 3.4: Video Labeling. AOC: Pedestrian (p).

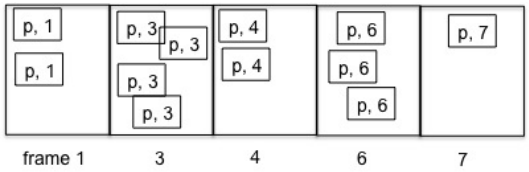

Figure 3.5: The Single Equivalent Frame Group for the Video Depicted in Figure 3.4

of the same color. Figure 3.7 depicts the equivalent frame groups for the video in Figure 3.6. Note that, by this definition, frame four qualifies as a member of equivalent frame groups one and two. As such, equivalent frame groups are not considered true equivalence classes.

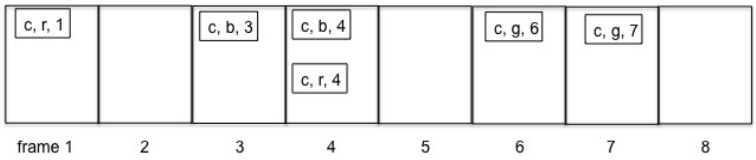

Figure 3.6: Video Labeling. AOC: Car (c). POC: Color: Red (r), Green (g), Blue (b).

Single-AOC, No POCs, AOC-Number CC, Global TC

Now, consider the summarization ideal where the AOC is pedestrian, no POCs are specified, the CC consists of the number of AOCs, that is, frames with different numbers of pedestrians are considered distinct, and the TC is global. In this case, two frames are equivalent if they have the same number of pedestrians. Figure 3.9 depicts the equivalent frame groups for the video in Figure 3.8.

Multiple-AOC, No POCs, AOC-Combination CC, Global TC

Now, consider the summarization ideal where the AOCs are car and pedestrian, no POCs are specified, the CC consists of the different AOC combinations, and the TC is global. Here, the possible AOC combinations are car, pedestrian, and car-pedestrian. Two frames are equivalent if they have the same combination of cars and pedestrians without regard to order or number. Once again, POCs, such as a pedestrian's identity, are not specified. This

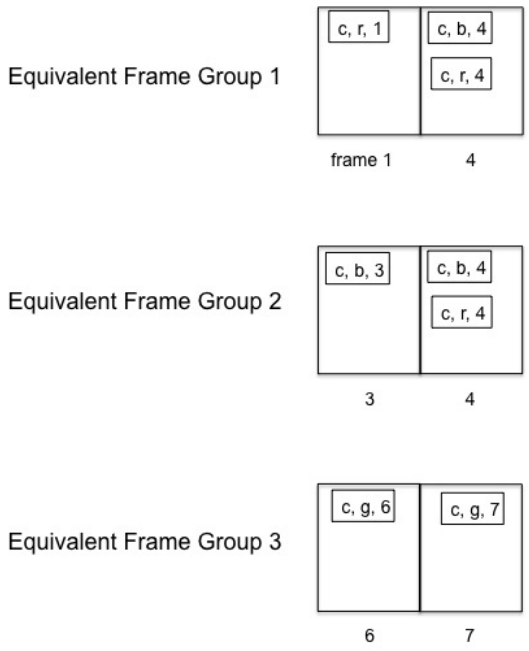

Figure 3.7: Equivalent Frame Groups for the Video Depicted in Figure 3.6

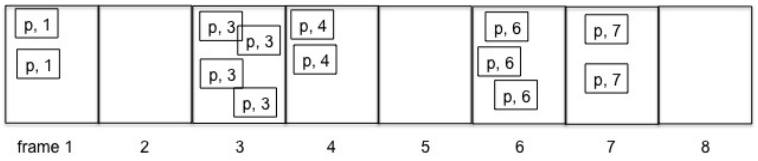

Figure 3.8: Video Labeling. AOC: Pedestrian (p).

means that two frames with pedestrians, and pedestrians only, even with different identities are equivalent. Figure 3.11 depicts the equivalent frame groups for the video in Figure 3.10.

Multiple-AOC, No POCs, AOC-Combination CC, Local TC

Finally, consider the summarization ideal where the AOCs are car and pedestrian, no POCs are specified, the CC consists of the different AOC combinations, and the TC is local. First, equivalent frames are computed for a global TC, as done in the previous section. Next, equivalent frame groups are formed by grouping consecutive equivalent frames (i.e., there can be no gap between equivalent frames). Figure 3.13 depicts the equivalent frame groups for the video in Figure 3.12.

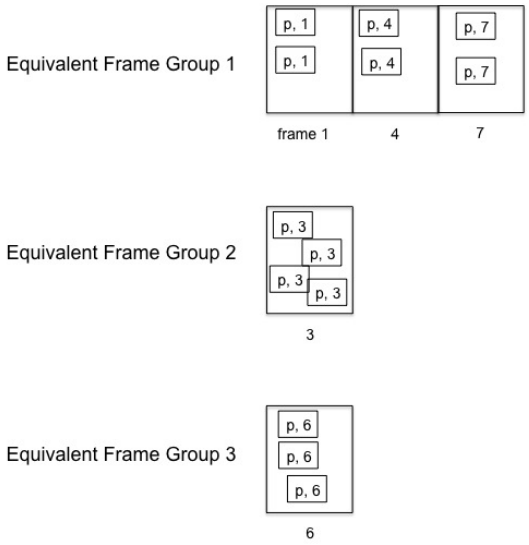

Figure 3.9: Equivalent Frame Groups for the Video Depicted in Figure 3.8

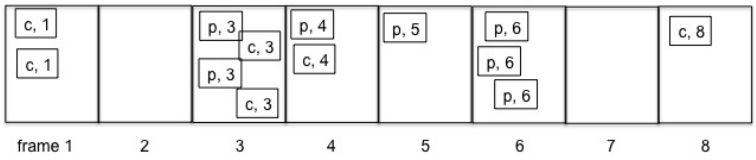

Figure 3.10: Video Labeling. AOCs: Pedestrian (p), Car (c).

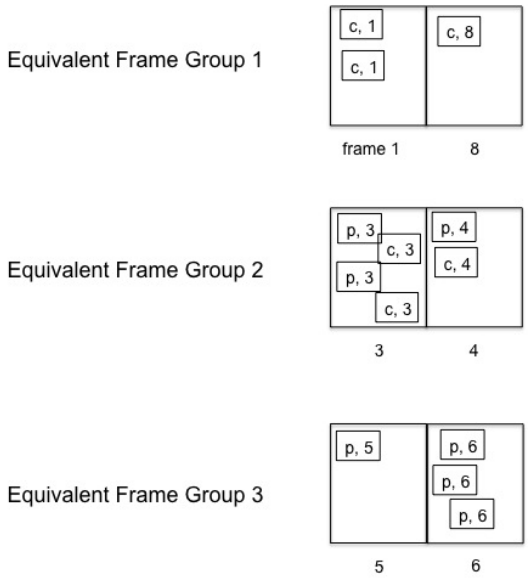

Figure 3.11: Equivalent Frame Groups for the Video Depicted in Figure 3.10

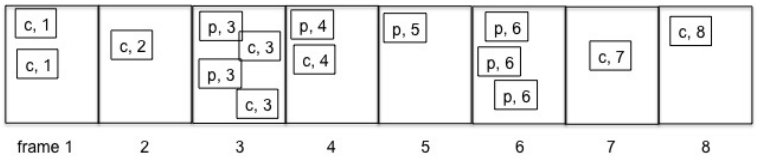

Figure 3.12: Video Labeling. AOCs: Pedestrian (p), Car (c).

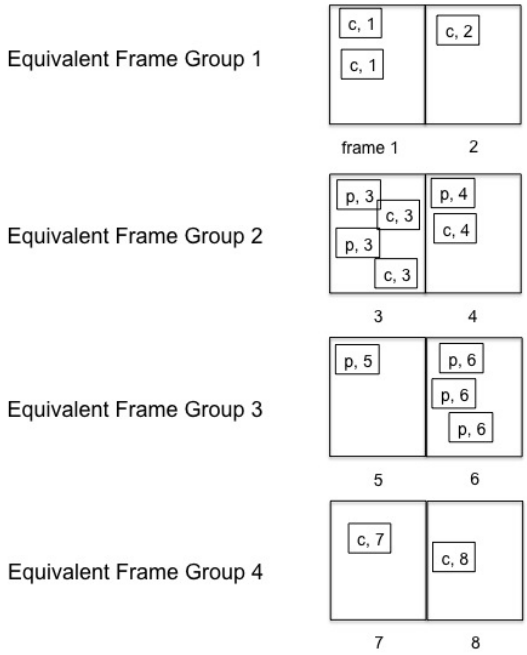

Figure 3.13: Equivalent Frame Groups for the Video Depicted in Figure 3.12

3.2.2 Summary Quality Metrics

Given the ground truth summary of a video, as determined above, the author proposes the following quantitative measures to compare it to a given keyframe summary. The results may be used to determine the keyframe summary's quality.

Variability

Sometimes, it may be desirable to measure the variability of a keyframe summary with respect to its ground truth summary. For example, consider a ground truth with seven equivalent frame groups. Also, consider three keyframe summaries, each of size 14 frames: A, B, and C. Summary A consists of one frame from each of the seven equivalent frame groups plus seven more from just a single group. Summary B, on the other hand, consists of seven frames from one equivalent frame group and seven frames from another group. Summary

C consists of two frames from each of the seven equivalent frame groups. Intuitively, summary C may be considered the most variable, and summary B, the least. The summary frame distributions for summaries A, B, and C are depicted in Figure 3.14, Figure 3.15, and Figure 3.16, respectively.

If keyframes possessed meaningful numerical values, then one could simply calculate a summary's (sample) variance just as one can calculate the variance in height associated with a sample of people drawn from the human population. However, each keyframe is assigned a nominal value, that is, an equivalent frame group. As such, the author quantifies the variability of a keyframe summary based uponstatistical dispersion measures for nominal distributions. Three measures are presented: variation ratio, information entropy, and variation around the mode. Variation around the mode is ultimately chosen because its value is between 0 and 1, but the author acknowledges that the other measures may convey keyframe summary variability equally well.

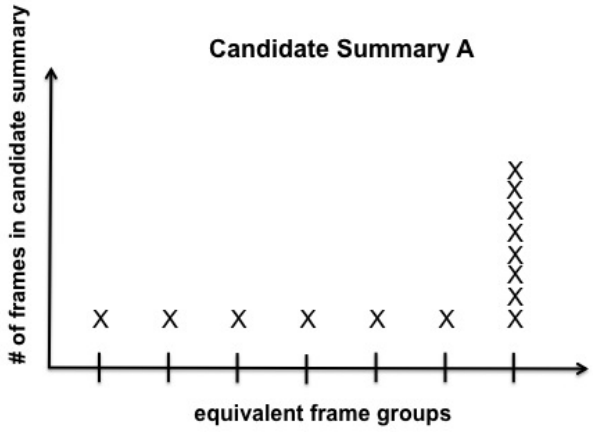

Figure 3.14: Keyframe Summary A Frame Distribution

Variation Ratio

keyframe summary variability may be defined as the variation ratio, v_r, where

$$v_r = 1 - \frac{f_m}{|\mathbf{K}|}$$

Where f_m denotes the frequency of the mode. Consider summary A. The mode is defined as the equivalent frame group with the most keyframes, and the frequency of the mode, in this case, is 8 ($f_m = 8$). $\mathbf{K}$ denotes the set of keyframes, so $|\mathbf{K}|$, in our example, is 14.

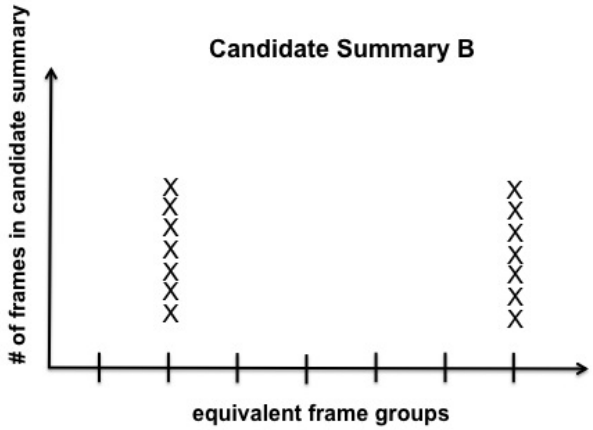

Figure 3.15: Keyframe Summary B Frame Distribution

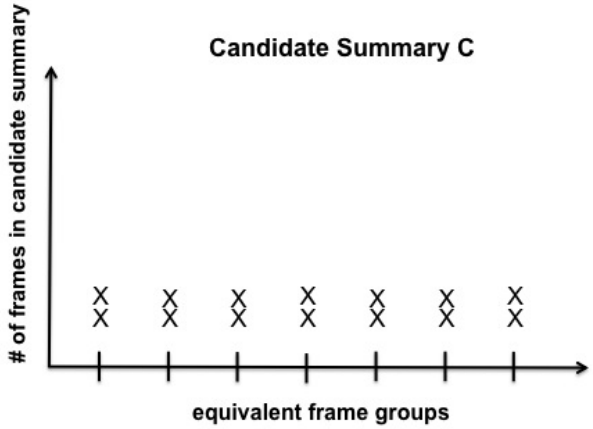

Figure 3.16: Keyframe Summary C Frame Distribution

Therefore, the $v_r = 1 - \frac{8}{14} = 0.43$. For summary C, however, the $v_r = 1 - \frac{2}{14} = 0.86$. C is more variable than A [50].

Wilcox's Indices

keyframe summary variability may also be defined as the variation around the mode, v_m, where

$$v_m = \frac{|\mathbf{E}| \times v_r}{|\mathbf{E}| - 1}$$

Here, $\mathbf{E}$ denotes the set of equivalent frame groups, so $|\mathbf{E}|$, in our example, is seven. v_r denotes the variation ratio, as defined above. The v_m for summary C, then, is 1, which is

the highest possible value and indicates a completely variable summary [51].

Information Entropy

keyframe summary variability may also be defined as the information entropy, i_e. In the context of the above example,

$$i_e = -\sum_i^E p(i) \log_2 p(i)$$

where $p(i)$ is the ratio of the number of keyframes in equivalent frame group i to the total number of keyframes. Under this scheme, summary C has an $i_e = 1.95$ and summary A has an $i_e = 1.45$. In the above example, if all of the frames were concentrated in a single equivalent frame group, then the resulting i_e would be 0, which is the lowest possible variability score. There is no upper limit on information entropy [52].

Sensitivity and False Positive Rate

Certain equivalent frame groups within a ground truth summary are selected by a keyframe summary and others may be missed. This can be viewed as a binary classification problem on the set of video frames. A keyframe summary's quality may be computed with standard binary classification metrics. The author chooses sensitivity, s_n, and false positive rate, f_r, the definitions of which are reviewed below.

$$s_n = \frac{\#t_p}{\#t_p + \#f_n}$$

$$f_r = \frac{\#f_p}{\#f_p + \#t_n}$$

A true positive, t_p, is a frame in the keyframe summary that belongs to an equivalent frame group that has not been selected already. A false positive, f_p, is a frame in the keyframe summary that does not belong to an equivalent frame group or belongs to an equivalent frame group that has already been selected. A true negative, t_n, is a frame that is absent from the keyframe summary, which does not belong to an equivalent frame group or belongs to an equivalent frame group that has already been selected. A false negative, f_n, is a frame that is absent from the keyframe summary, which belongs to an equivalent frame group that has not been selected already.

The ground truth summary determines which keyframes should have and should not have been selected. Consider a ground truth summary with n equivalent frame groups. A keyframe summary's sensitivity and false positive rate can be computed as follows.

Let $\mathbf{K}_i$ be the set of keyframes in equivalent frame group i, so $\mathbf{K}_i \subseteq \mathbf{K}$ and $\bigcap_i \mathbf{K}_i = \mathbf{K}$. Here, i is an integer, ranging from 1 to n. Also, let

$$\mathbf{f}(x) = \begin{cases} 1 & \text{if } x \geq 1, \\ 0 & \text{otherwise} \end{cases}$$

then,

$\#t_p = \sum_{i=1}^{n} \mathbf{f}(|\mathbf{K}_i|)$, or the number of equivalent frame groups from which at least one frame is in the keyframe summary,

$\#t_p + \#f_n = n$,

$\#f_p = |\mathbf{K}| - \#t_p$, where $\mathbf{K}$ is the set of keyframes (note: this considers all but exactly one frame per equivalent frame group as a f_p),

$\#f_p + \#t_n = |\mathbf{V}| - n$, where $\mathbf{V}$ is the set of video frames.

Sensitivity and false positive rate are chosen for intuitive reasons. Sensitivity can be seen as a measure of the extent to which the keyframe summary contains all of the ground truth content. False positive rate, on the other hand, can be seen as a measure of the amount of redundancy in a keyframe summary.

Receiver Operating Characteristic (ROC) Curve

In this work a particular summarization method is better than another method if it has better sensitivity for any given false positive rate. More precisely, one method is better than another method if the area under the curve (AUC) of its respective ROC curve is larger. Here, the x and y-axes of the ROC coordinate system are designated as the false positive rate and sensitivity, respectively.

An ROC curve for a particular summarization system may be computed as follows. First, produce summaries of various sizes, the interval and range of which depend on summary stakeholders' goals, among other factors. For each summary that is produced, compute its false positive rate and sensitivity, as outlined in the previous section, and plot it on the ROC coordinate system. The points are then connected to construct the final "curve," and the AUC can be computed.

Of note, the overall AUC of one summarization method may be smaller than that of another, but it may possess a larger AUC on a particular subset of x-axis ranges. If these particular x-axis ranges are of interest to the summary stakeholders, then the overall AUC score may be misleading. As such, the ROC curves should be viewed in concert with the overall AUC score.

A stakeholder may not be interested in the performance of a summarization method on a single video, but rather a collection of videos, perhaps of the same type. The performance on the collection may provide a better indication of average algorithm performance. So, instead of computing an ROC curve for an individual video, a combined ROC curve for a video collection may be computed. This can be accomplished by averaging the sensitivities for a given false positive rate for all of the videos in the collection. AUC computation remains the same.

Dissimilarity

Another measure of keyframe summary quality is how much the distribution of keyframes differs from the distribution of ground truth frames. For example, consider a ground truth summary with three equivalent frame groups: A, B, and C. Let us also say that equivalent frame group A consists of 70 frames, B consists of 20 frames, and C consists of 10 frames. Keyframes are distributed exactly as the ground truth if 70 percent of the keyframes come from group A, 20 percent from group B, and 10 percent from group C.

The ground truth summary histogram may be constructed as follows. The x-axis units represent the different equivalent frame groups of a video. The y-axis value, for a respective equivalent frame group, is the proportion of frames in that group over the total number of frames in the ground truth summary. The keyframe summary histogram may be computed in a similar fashion. The x-axis units still represent the different equivalent frame groups

of the video. The y-axis value, for a respective equivalent frame group, is equal to the proportion of keyframes in that group over the total number of keyframes. Example ground truth and keyframe summary histograms are depicted in Figure 3.17 for a ground truth summary with three equivalent frame groups. The keyframe summary histogram is in grey and the ground truth summary histogram is in black.

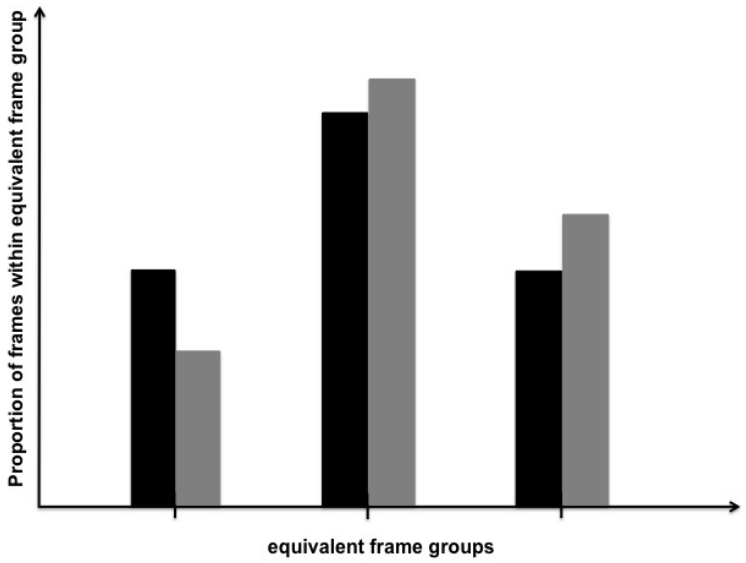

Figure 3.17: Ground Truth and Keyframe Summary Histograms

This work defines a keyframe summary's dissimilarity, d_s, as the city block distance between the ground truth summary and keyframe summary histogram. Other histogram difference measures exist, such as Euclidean and Bhattacharyya distances. City block distance is chosen for its simplicity. More formally,

$$d_s = \sum_i^E \ |\mathrm{H}_G(\mathrm{i}) - \mathrm{H}_C(\mathrm{i})|,$$

where **E** is the set of equivalent frame groups, H_G is the ground truth summary histogram, and H_C is the keyframe summary histogram. A larger score indicates that the keyframe summary is more dissimilar to the ground truth summary.

CHAPTER 4:
Summarization Process Overview

This work now examines the summarization process as a whole and how the author's evaluation framework has been incorporated. The ultimate goal of the process is to produce keyframe summary evaluation results. The results can be used to assess the quality of keyframe summaries and the methods that produced them.

Different stakeholders may focus on different subprocesses of the overall summarization process. For example, the keyframe summary consumer may be especially interested in ensuring their summarization concerns are properly captured by the summarization ideal. The keyframe summary algorithm developer, on the other hand, may play a limited role in the specification on the summarization ideal. Instead, their focus will be directed toward selecting and developing the appropriate technologies that satisfy the consumer's concerns.

An outline of the summarization process is provided in Figure 4.1. The remaining sections expand upon the processes displayed here. In the follow-on chapter, the overall process will be "executed" to accomplish a number of goals.

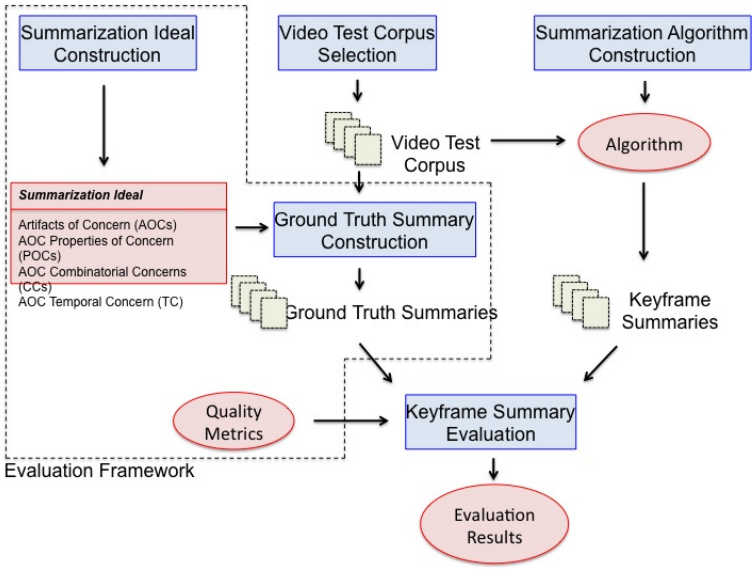

Figure 4.1: Summarization Process

4.1 Summarization Ideal Construction

Summarization ideal construction is the process of specifying the AOCs, POCs, CCs, and TCs. The process may not be automated. It requires stakeholders to consider their summarization needs and translate them into a summarization ideal, as defined in Section 3.1.

4.2 Video Test Corpus Selection

Video test corpus selection entails collecting a number of videos, the number and type of which may depend on the summarization stakeholders' needs. For example, a stakeholder interested in baseball artifacts, such as pitches, catches, and swings, may select videos from the baseball genre. A larger number of videos in the test corpus typically result in a better understanding of expected algorithm performance.

4.3 Ground Truth Summary Construction

Ground truth summary construction is depicted in Figure 4.2. For each video in the test corpus, the AOCs are manually labeled and their respective POC values. Based on the summarization ideal (AOCs, POCs, CCs, and TCs), equivalent frames are then automatically computed and grouped to form the ground truth summary. Ground truth summary construction is discussed in detail in Section 3.2.1.

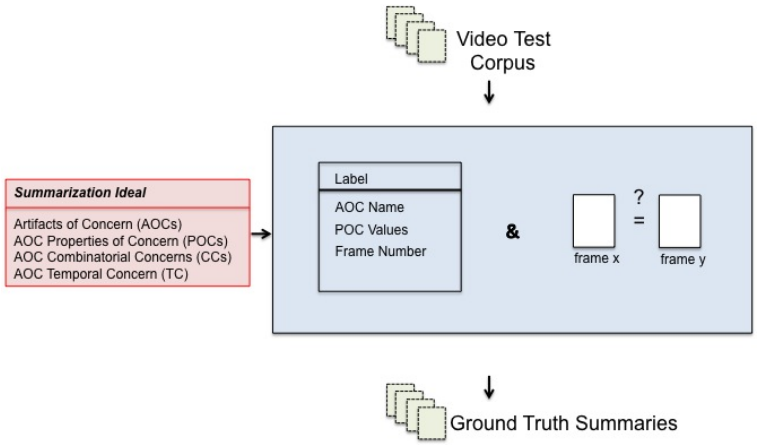

Figure 4.2: Ground Truth Summary Construction

4.4 Summarization Algorithm Development

Summarization algorithm development is depicted in Figure 4.3. This is the most creative component of the summarization process, and is the target of the evaluation framework. The process of algorithm construction may not be rigorously defined, but, in general, it consists of the application of summarization technologies to satisfy the stakeholder's summarization ideal. The stakeholder's summarization ideal, then, serves as a guide in this process. For example, if the stakeholer's AOC is "face," then facial tracking technology may be particularly relevant. The nature of the video test corpus may also serve as a guide. For example, if the test videos possess clear shot boundaries, then SBD technology may be employed effectively to segment the videos into their shots. An algorithm is deemed to be good if, in the end, it produces good keyframe summaries.

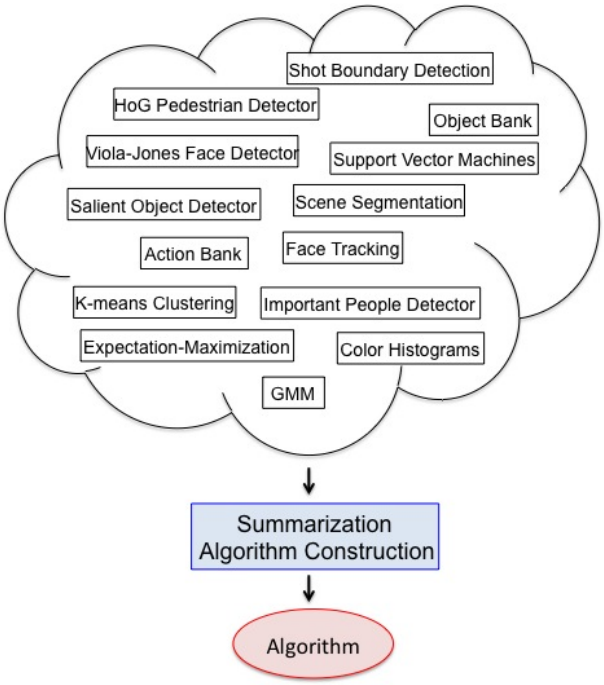

Figure 4.3: Summarization Algorithm Development

4.5 Keyframe Summary Evaluation

Keyframe summary evaluation is depicted in Figure 4.4. For each of the summaries produced by a summarization algorithm, the metrics of Section 3.2.2 are computed. The resulting scores, then, are used to evaluate how well the summarization algorithm performed.

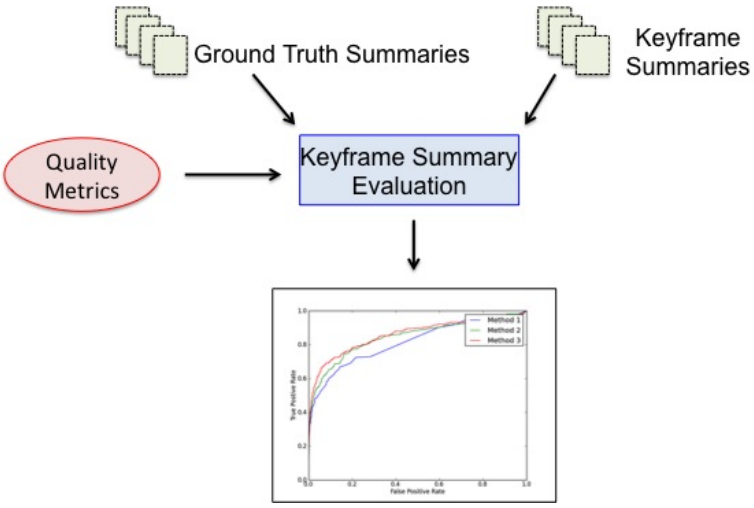

Figure 4.4: Keyframe Summary Evaluation

CHAPTER 5:
Summarization Process Examples

The previous chapter described, in general terms, the summarization process followed in this paper. This chapter provides concrete examples of the summarization process for greater clarity. The follow-on chapter examines the output of the summarization process: the keyframe summary evaluation results.

In this chapter three different summarization ideals are constructed. For each summarization ideal, test corpus videos are selected and their ground truth summaries are created. A number of summarization algorithms are developed to satisfy the summarization ideals.

5.1 Summarization Ideal Construction

The author assumes the role of summary stakeholder and defines summarization ideals centered around three AOCs: scenery types, faces, and important people.

5.1.1 Scenery Types

5 scenery types are selected as AOCs. They are indoor, outdoor, natural, man-made, and, greenery. More than one scenery type may exist in a given frame. For example, an image of a lake surrounded by a forest may be classified as outdoor, natural, and greenery. The stakeholder does not specify any POCs. The stakeholder is interested in capturing the scenery content between scenery transitions. If a video transitions from indoor to outdoor, and then back to indoor, for example, then content from both indoor portions and the middle, outdoor portion is desired. This stakeholder desire is interpreted as an AOC-combination CC and a local TC. The resulting summarization ideal is depicted in Figure 5.1.

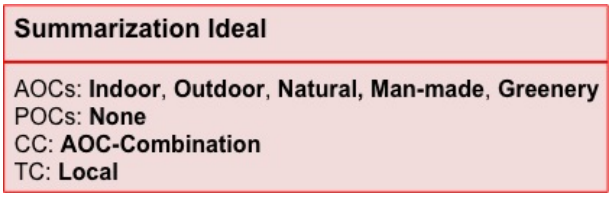

Figure 5.1: Scenery-Centered Summarization Ideal

5.1.2 Faces

The human face is selected as the single AOC. The stakeholder considers a human face to be present when two eyes, a nose and mouth are visible. A face is also considered to be present if the subject's eyes are closed or is wearing (sun)glasses. The stakeholder is not interested in any face properties, such as identity or race. The stakeholder has no face combinatorial concerns. A frame with one face is equally important as a frame with two or three faces, for example. However, the stakeholder is interested in capturing content from the temporally different portions of a video. This stakeholder desire is interpreted as an local TC. The resulting summarization ideal is depicted in Figure 5.2.

Figure 5.2: Face-Centered Summarization Ideal

5.1.3 Important People

Important persons are selected as our AOC. An important person, like the protagonist in a narrative work, is a primary subject of interest. The stakeholder considers an important person to be present if half of their body or face is visible. They are interested in capturing all the important persons, in their different environments, in different size groups. Identity and place POCs are added to the stakeholder's summarization ideal to accommodate these concerns. The stakeholder also has an AOC-number CC, that is, a portion of video with two important people is distinct from a portion of video with three important people. Finally, the stakeholder has a Global TC, that is, temporally different portions of a video with the same important people in the same place are not considered distinct. The resulting summarization ideal is depicted in Figure 5.3.

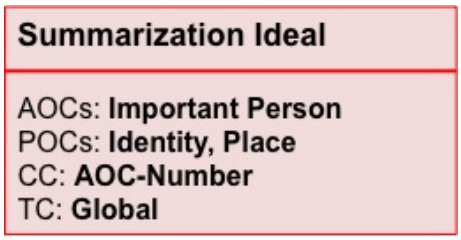

Figure 5.3: Important-Person-Centered Summarization Ideal

5.2 Video Corpus Selection

5.2.1 Scenery Types

11 user-generated videos are manually collected from YouTube [53]. The videos vary in duration, but fall between one and four minutes, with a corpus average of 2 minutes and 25 seconds. Each video possesses all scenery types, along with numerous scenery type transitions. Three videos are filmed with camera phones, and the other eight are filmed with standalone digital cameras. In terms of editing effects, six videos employ hard cuts to transition between shots. The other five videos consist of continuous video with no distinct shot transitions.

5.2.2 Faces

10 user-generated videos are manually collected from YouTube [53]. The videos vary in duration, but fall between one and six minutes, with a corpus average of 3 minutes and 5 seconds. Each video possesses a number of human faces. All videos possess "face-present" and "face-absent" intervals of various length and number. Additionally, most of the videos contain intervals in which more than one face is present. Three videos are filmed with Google Glass. It is unclear what devices the remaining video were filmed with. In terms of editing effects, all of the videos employ hard cuts.

5.2.3 Important People

10 user-generated videos are manually collected from YouTube [53], which exhibit various important people in various places over the course of the video. Again, an important person is defined as a primary subject of interest in an overarching narrative. Similar to the other video collections, the average video is short (a few minutes), and is filmed with various camera technologies.

Future work may consider increasing the respective corpus sizes to strengthen the results.

5.3 Ground Truth Summary Construction

For each summarization ideal presented above, this work employs the Video Annotation Tool from Irvine California (VATIC) to manually label each AOC of each frame of each video [54]. The video labeling, combined with the definition of frame equivalency, sufficiently specify the ground truth summary for a video.

The AOC labeling was specified in Section 3.2.1 and is straightforward. Various definitions for frame equivalency were covered in Section 3.2.1, but greater clarity is provided for each summarization ideal below.

5.3.1 Scenery Types

This work utilizes the frame equivalency definition provided in Section 3.2.1 for a Multiple-AOC, No POCs, AOC-Combination CC, Local TC summarization ideal. Two frames are equivalent if the same scenery types are present in the same combination without regard to order or number. Also, the frames cannot be separated by a single non-equivalent frame.

5.3.2 Faces

Two frames are equivalent if they have a face present. Also, the frames cannot be separated by a frame in which no face is present. Otherwise, they are considered different.

It is clear that temporal equality may be defined in a different way. For example, perhaps one could allow the gap between frames with faces to be greater than a single frame for them to be considered equal.

5.3.3 Important People

Two frames are equivalent if they have the same important people in the same place. Two places are equal if they are the same indoor room. If the places are outdoors, then two places are equal if they are in the same general area.

5.4 Summarization Algorithms

The author developed two baseline summarization methods along with a number of other, more sophisticated, content-dependent algorithms. The baseline summarization methods are the same for each summarization ideal.

5.4.1 Uniform Random Frame Sample

Consider the discrete uniform distribution, *U(1,f)*, where f is the number of frames in a given video. For a keyframe summary of size s, the first baseline method randomly selects s integers from the distribution without replacement. The video frames with these values represent the final keyframe summary.

5.4.2 Evenly-Spaced Frame Sample

The second baseline method consists of selecting keyframes at approximately evenly-spaced intervals from the target video. Evenly spaced frame numbers are computed, but frame number that is selected is perturbed by ± 5 percent of the total number of video frames. These approaches may be classified as content-independent, that is, keyframes are not selected based on video content.

The summarization ideals developed in this chapter are semantically high-level, that is, the summary stakeholder has specified high-level AOCs (scenery types, faces, and important people), and other high-level concerns. It is hypothesized that, when it comes to these types of summarization ideals, better summaries can be constructed than the baseline approaches by employing high-level, albeit imperfect, content detectors.

5.4.3 Scenery Types

Scenery Transition Peaks

5 separate binary SVM classifiers (one for each scenery type) are constructed as follows. For each scenery type, the author gathered a training set of images with an equal number of positive and negative examples. Training examples are collected from Oliva and Torralba, Quattoni and Torralba, and ImageNet [45, 55, 56].

This work represents each training image as a 52,134-dimensional object bank feature vector [43]. A Library for Support Vector Machines (LIBSVM) is then used to construct the

respective binary SVM classifiers [57].

Each frame of a given video is classified according to each of the five binary classifiers. The result is visualized as in Figure 5.5, where the x-axis is the frame number and colored segments represent positive scenery classifications.

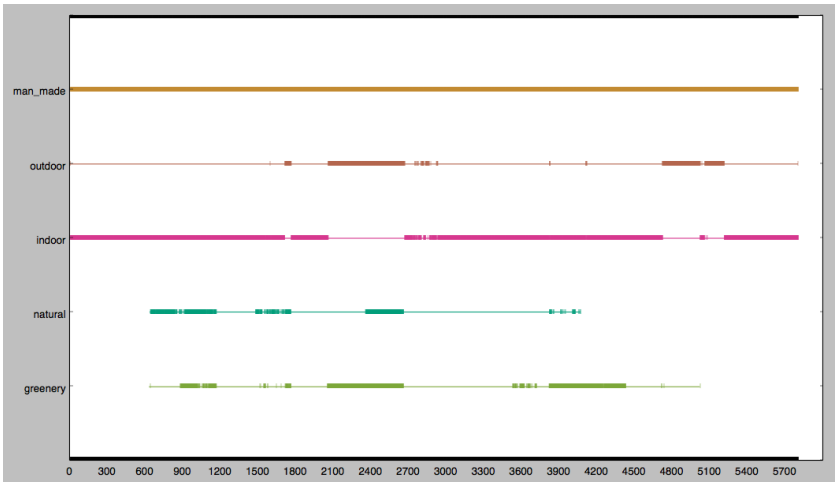

Figure 5.4: Per-Frame Scenery Classification

The number of scenery transitions from one frame to the next is then counted. The counts are then smoothed with a Hamming window of 121 frames, which equates to about four seconds for a video produced at 30 frames per second. The result is visualized as in Figure 5.5, where the x-axis is the frame number and the y-axis is the smoothed count of scenery transitions from frame n to frame $n+1$.

Local maxima, or transition peaks, of the smoothed counts are identified by convolving the smoothed count signal with a continuous wavelet transform with a window size of 31 frames. Peaks that appear within this window are then selected. Keyframes from the segments defined by the transition peaks are then selected. Frames are selected in the following priority order. Middle frames of the largest segments have the highest priority. If the summary size is large enough such that all of the middle frames are selected, then the author cycles through the segments and picks frames at random.

The intuition for this algorithm is as follows. If the scenery transitions are perfectly identified, then this method could satisfy the summarization ideal perfectly by selecting content

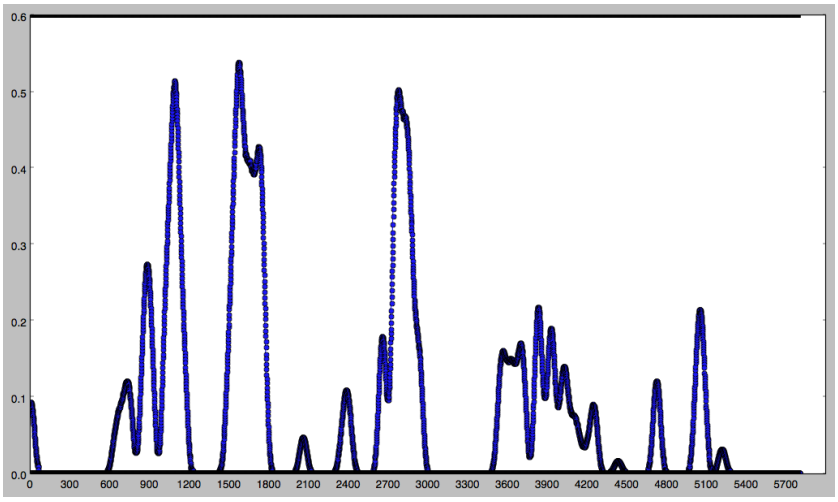

Figure 5.5: Inter-Frame Scenery Transition Count (Smoothed)

from each segment.

5.4.4 Faces

Face Segments

Using the Viola-Jones face detector provided in the Easy! Computer Vision API, this method classifies each frame of a given video as possessing a face or not [58]. The method then smooths this signal with a Hamming window of 61 frames. The result is visualized as in Figure 5.6. The x-axis represents the frame number and the black segments are portions of the video where a face is predicted. A segment is simply a continuous sequence of positive face detections and, in this example, there are 10 segments. Finally, for a given summary size, the method cycles through the segments, each time selecting a frame at random, without replacement.

Intuitively, if the method identifies the face segments perfectly, then it could satisfy the summarization ideal perfectly by selecting content from each segment.

Face Peaks

First, this method predicts the face segments as in the "Face Segments" method. However, it also keeps track of the number of faces detected in each frame. Next, like in the "Scenery Transition Peaks" method, this method determine the peaks in the number of face transitions. If a peak occurs within a face segment, then the segment is divided into separate

Figure 5.6: Predicted Video Face Segments (Smoothed)

portions delineated by this frame. The method then cycles through these new segments, each time selecting a frame at random, without replacement.

Intuitively, this algorithm may be more appropriate when an AOC-count CC is specified in the summarization ideal, but it is evaluated none-the-less.

Face Frames Color Histogram Clustering

Again, using the Viola-Jones face detector provided in the Easy! Computer Vision API, this method classifies each frame of a given video as possessing a face or not [58]. It represents each positively classified frame as an RGB color histogram. This method clusters all of the frames using k-means (k equals the summary size), with a Euclidean distance metric. Finally, the frame closest to each cluster's center is selected for summary inclusion.

Intuitively, the different portions of a video with faces (i.e., face segments) might possess similar color histograms. Consequently, each cluster of positively classified frames might belong to the same face segment. Picking a frame from each cluster could potentially satisfy the summarization ideal.

5.4.5 Important People

Salient Objects

First, for each frame of a given video, this method computes the frame's saliency map using technology presented by Jiang et al. [37]. The saliency map is a grayscale representation of the original image. Each pixel is assigned an intensity value that corresponds to its saliency score. Pixels with greater saliency receive higher intensity values. Next, for each frame, the method computes the color histogram. However, it only considers pixels whose corresponding saliency map values exceed a given threshold. In essence, it computes the RGB color histogram of the salient objects. This method clusters all of the frames using

k-means (k equals the summary size), with a Euclidean distance metric. Finally, the frame closest to each cluster's center is selected for summary inclusion.

Although not a rule, important people, or the primary subjects of a video's narrative, often occupy prominent positions within a frame. As a result, the salient object detector may capture the important people. Of course, the detector may capture salient objects other than people and it may miss important people altogether. Intuitively, important people in different places may possess significantly different color histograms, so, selecting frames from the different clusters may satisfy the stakeholder's POCs. Also, let us say that a frame contains person A and another frame contains person A and person B in the same place. By employing color histograms of the salient objects, this technique may place the two frames in different clusters. Hence, the stakeholder's AOC-number CC may be satisfied.

Face and Pedestrian Detectors

Using the Viola-Jones face detector and the HOG pedestrian detector provided in the Easy! Computer Vision API, this method classifies each frame of a given video as possessing a face or pedestrian [58]. Keyframes are then selected from the positively classified frames in the same manner as in the "Face Frames Color Histogram Clustering" algorithm presented in Section 5.4.4, which is based on frame color histograms.

Intuitively, the face and pedestrian detectors may detect a portion of the important people within a video. Of course, not every face or pedestrian they detect will be important. They may also miss faces and pedestrians or assign positive classifications to non-faces and non-pedestrians. Regardless, using the same rationale as the previous approach, clustering the frames with positive detections on their color histograms may allow us to satisfy the stakeholder's summarization ideal. Additionally, clustering on the color histogram of the entire frame, instead of just the face or important person, may help satisfy the place POC to a greater extent than the salient object detector approach.

Combined

This method combines the two algorithms presented above. First, this method identifies positive frames as in the "Face and Pedestrian Detectors" algorithms. Then, for the positive frames, it applies saliency maps and selects keyframes as in the "Salient Objects" algorithm.

The method attempts to minimize the false detections made by the salient object detector by presenting it with a potentially higher concentration of frames with pedestrians and faces.

CHAPTER 6:
Results

This chapter presents the output of the summarization process: the keyframe summary evaluation results. The results are simply presented here and the follow-on chapter discusses how they support the achievement of the author's thesis goals.

6.1 Methodology

The author executes each summarization algorithm developed in the previous chapter to produce keyframe summaries of different sizes. The algorithms are only executed on the test corpus videos for which the algorithms were designed. For example, the "scenery type" algorithms are only executed on the scenery corpus videos.

The author generates summaries at intervals of 0.005 percent of a given video's total size with each algorithm. For example, if a test corpus video is 2000 frames long, then summaries of size one frame (0.0005×2000), two frames (0.001×2000), ..., and 40 frames (0.02×2000) are generated. Summary sizes are limited to two percent or less of the total video size because the performance of the algorithms under consideration converge for larger summaries, that is, the algorithms perform equally well for summaries larger than about two percent.

Consider points one, two, and three in Figure 6.1. For a false positive rate of 0.005, the "Scenery Transition Peaks" method achieves the best sensitivity. More specifically, the "Scenery Transition Peaks" method generated keyframe summaries at "$0.0005 \times$video size" intervals for each of the 11 scenery test videos. The false positive rate and sensitivity was computed for each of these summaries. The average sensitivity was then computed for all of the summaries with a 0.005 false positive rate and then plotted (point one). Point one has a greater sensitivity than points two and three. This means that for a given level of redundancy, the "Scenery Transition Peaks" method produces keyframe summaries with more desirable content.

Consider points one, two, and three in Figure 6.2. For a keyframe summaries of size

$0.005 \times$ video size, the "Evenly-Spaced" method achieves the best variability. More specifically, the "Evenly-Spaced" method generated keyframe summaries at "$0.0005 \times$ video size" intervals for each of the 11 scenery test videos. The variability was computed for each of these summaries. The average variability was then computed for all of the summaries with a size of $0.005 \times$ video size (point one). Point one has a greater variability than points two and three. This means that for a summary size of $0.005 \times$ video size, the "Evenly-Spaced" method produces keyframe summaries that are the most variable, with respect to equivalent frame group membership. The dissimilarity curves are computed in the same fashion.

The AUCs for each curve are computed and presented in tabular form. For the ROC and variabilty curves, the highest and lowest possible AUC scores, in general, are 0.02 and 0.0, respectively (reversed for dissimilarity metric). Given a particular video, on the other hand, the highest and lowest possible AUC scores vary. For example, consider a video with 1000 frames in total and they all belong to a single equivalent frame group. Simply picking one frame captures 100 percent of the desired content. As such, algorithm scores are, in part, dependent upon the distribution of equivalent frames in the ground truth representations of the respective test videos. The algorithms are subjected to the same ground truth test videos, so the comparison is fair.

Finally, for each summarization ideal, visual results are presented. The author selects one video from the respective test video corpus. A keyframe summary of the video is produced using the best content-dependent algorithm and the baseline algorithms. They are displayed for comparison purposes.

6.2 Scenery Types

The AUC results for the "scenery type" algorithms presented in Section 5.4.3 are given in Table 6.1. Metric scores in bold font indicate the best score. The "Scenery Transition Peaks" algorithm achieved the best AUC for the ROC metric. The "Evenly-Spaced" algorithm achieved the best AUC for the variability and dissimilarity metrics.

The ROC, variability, and dissimilarity curves are presented in Figure 6.1, Figure 6.2, and Figure 6.3, respectively.

Figure 6.4, Figure 6.5, and Figure 6.6 display 15-frame keyframe summaries produced

	Quality Metrics		
Method	*ROC*	*Variability*	*Dissimilarity*
Uniform	0.01235	0.01172	0.006727
Evenly-Spaced	0.01252	**0.01229**	**0.004605**
Scenery Transition Peaks	**0.01332**	0.01180	0.007930

Table 6.1: Quality Metric AUCs (Scenery-Centered)

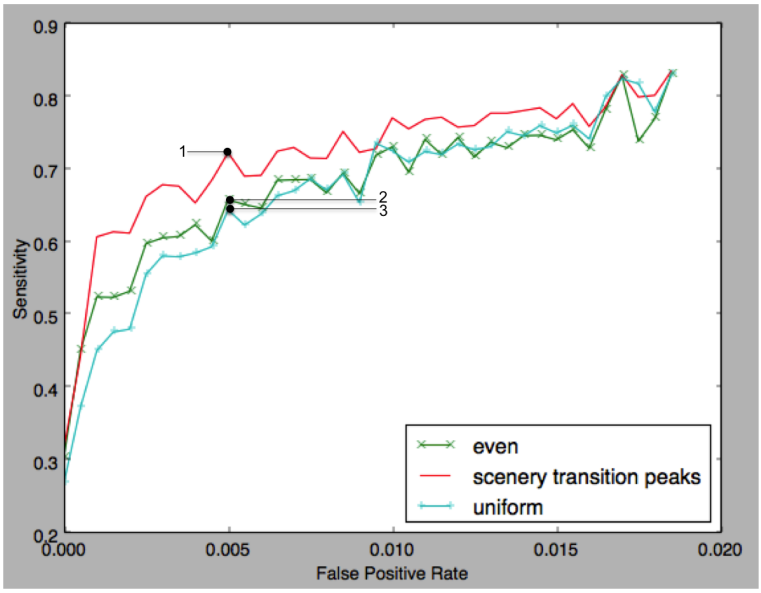

Figure 6.1: ROC Curves (Scenery-Centered, Larger Sensitivity is Better)

on a "scenery type" test corpus video by the "Uniform," "Evenly-spaced," and "Scenery Transition Peaks" methods, respectively. The video is 3 minutes and 45 seconds in duration.

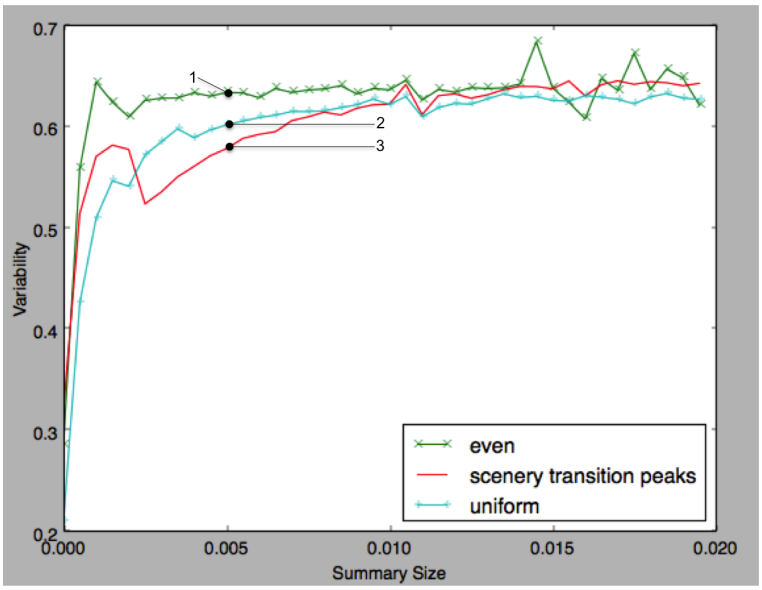

Figure 6.2: Variability Curves (Scenery-Centered, Larger Variability is Better)

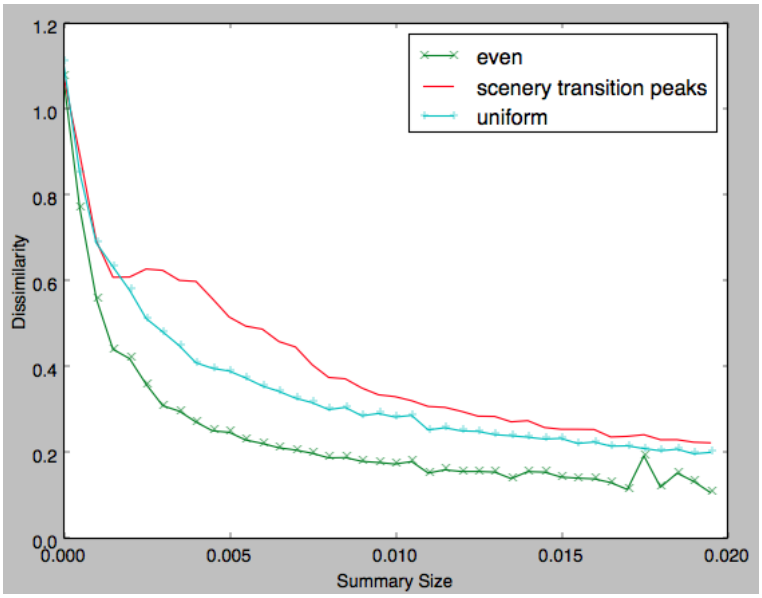

Figure 6.3: Dissimilarity Curves (Scenery-Centered, Larger Dissimilarity is Worse)

Figure 6.4: 15-Frame Keyframe Summary: "Uniform" Method

Figure 6.5: 15-Frame Keyframe Summary: "Evenly-Spaced" Method

Figure 6.6: 15-Frame Keyframe Summary: "Scenery Transition Peaks" Method

6.3 Faces

The AUC results for the "face" algorithms presented in Section 5.4.4 are presented in Table 6.2. The "Face Frames Color Histogram" algorithm achieved the best AUC for the ROC and dissimilarity metrics. The "Evenly-Spaced" algorithm achieved the best AUC for the variability metric.

The ROC, variability, and dissimilarity curves are presented in Figure 6.7, Figure 6.8, and Figure 6.9, respectively.

	Quality Metrics		
Method	*ROC*	*Variability*	*Dissimilarity*
Uniform	0.00865	0.01799	0.01570
Evenly-Spaced	0.00894	**0.01826**	0.01501
Face Segments	0.01008	0.01701	0.01615
Face Peaks	0.01026	0.01682	0.01580
Face Frames	**0.01196**	0.01634	**0.01478**

Table 6.2: Quality Metric AUCs (Face-Centered)

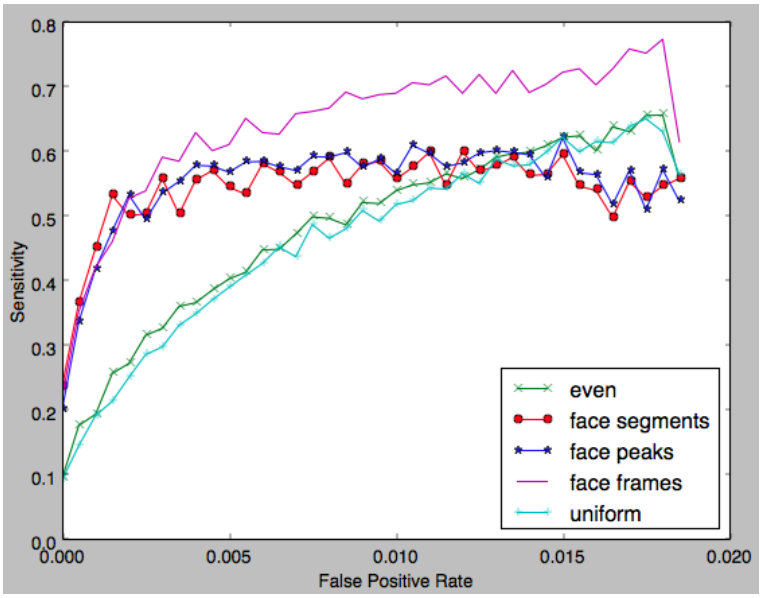

Figure 6.7: ROC Curves (Face-Centered)

Figure 6.10, Figure 6.11, and Figure 6.12 display 10-frame keyframe summaries produced on a "face" test corpus video by the "Uniform," "Evenly-spaced," and "Face Frames Color

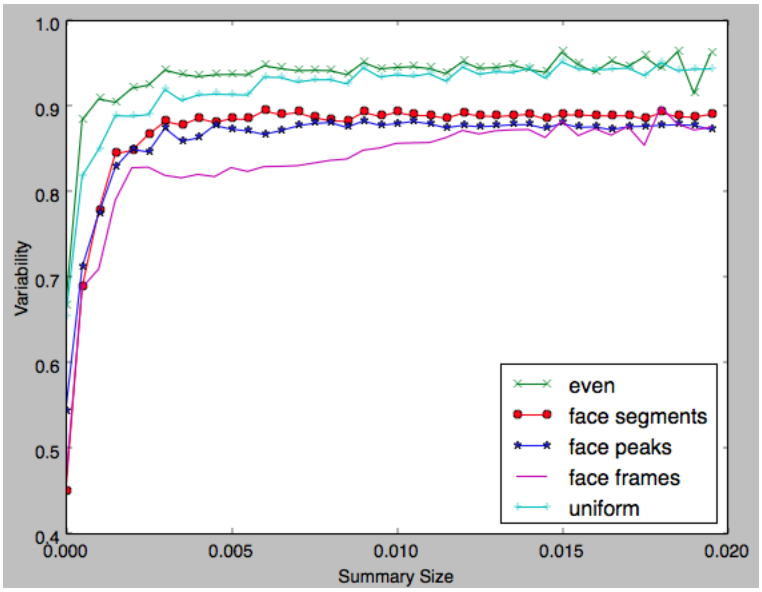

Figure 6.8: Variability Curves (Face-Centered)

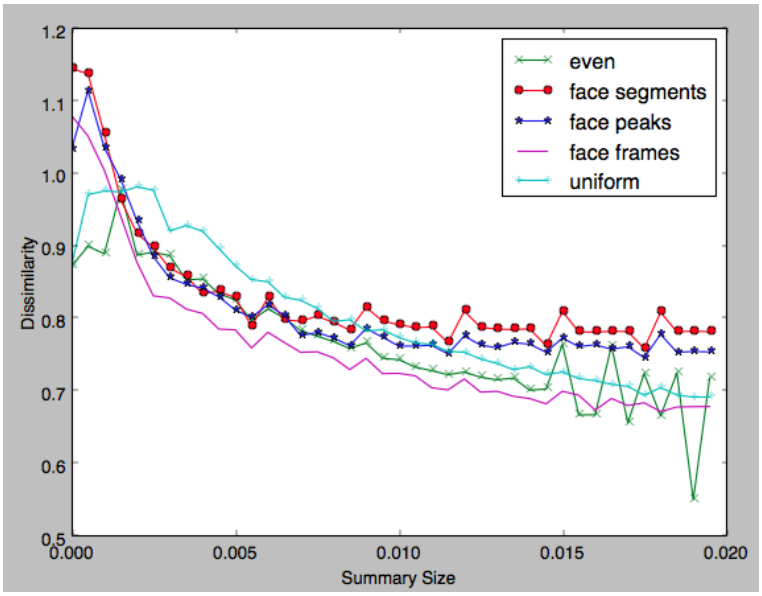

Figure 6.9: Dissimilarity Curves (Face-Centered)

Histogram Clustering" methods, respectively. The video is 3 minutes and 21 seconds in duration.

Figure 6.10: 10-Frame Keyframe Summary: "Uniform" Method

Figure 6.11: 10-Frame Keyframe Summary: "Evenly-Spaced" Method

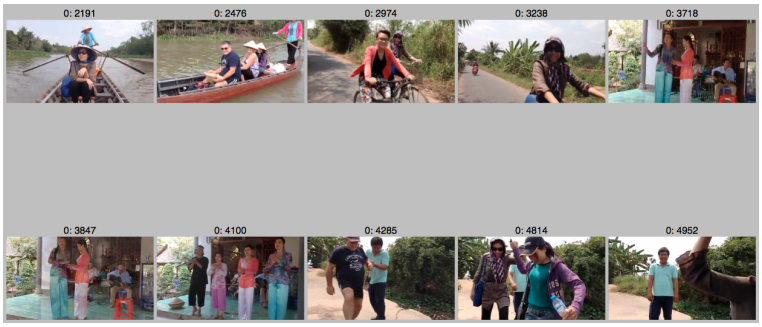

Figure 6.12: 10-Frame Keyframe Summary: "Face Frames Color Histogram Clustering" Method

6.4 Important People

The AUC results for the "important people" algorithms presented in Section 5.4.5 are presented in Table 6.3. The "Face and Pedestrian Detectors" algorithm achieved the best AUC for the ROC and variability metrics. The "Evenly-Spaced" algorithm achieved the best AUC for the dissimilarity metric.

The ROC, variability, and dissimilarity curves are presented in Figure 6.13, Figure 6.14,

and Figure 6.15, respectively.

	Quality Metrics		
Method	*ROC*	*Variability*	*Dissimilarity*
Uniform	0.01311	0.01461	0.01035
Evenly-Spaced	0.01350	0.01507	**0.00851**
Salient Objects	0.01337	0.01499	0.00997
Face and Pedestrian Detectors	**0.01507**	**0.01598**	0.00995
Combined	0.01409	0.01580	0.01089

Table 6.3: Quality Metric AUCs (Important-People-Centered)

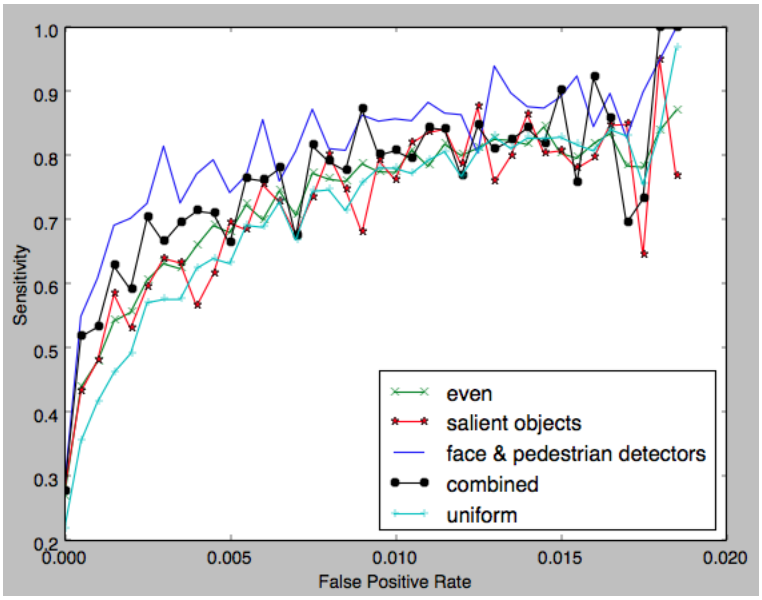

Figure 6.13: ROC Curves (Important-People-Centered)

Figure 6.16, Figure 6.17, and Figure 6.18 display 6-frame keyframe summaries produced on a "important people" test corpus video by the "Uniform," "Evenly-spaced," and "Face and Pedestrian Detectors" methods, respectively. The video is 3 minutes and 20 seconds in duration.

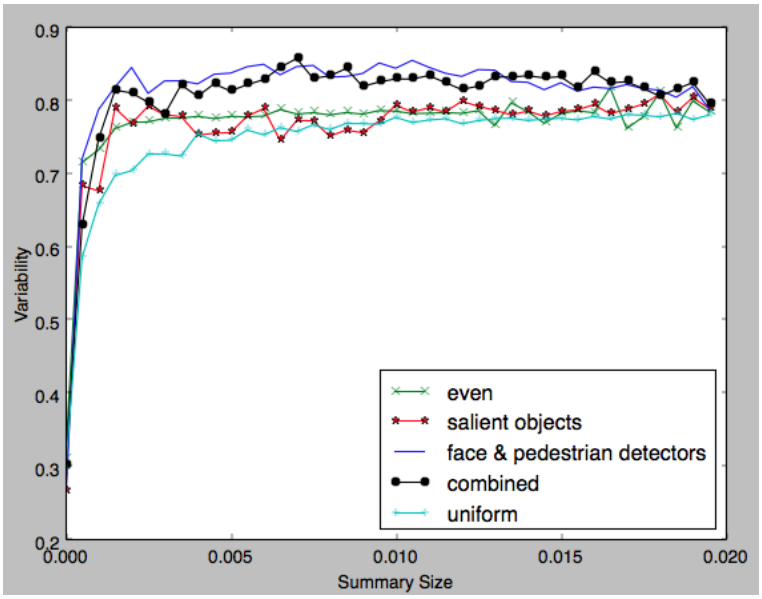

Figure 6.14: Variability Curves (Important-People-Centered)

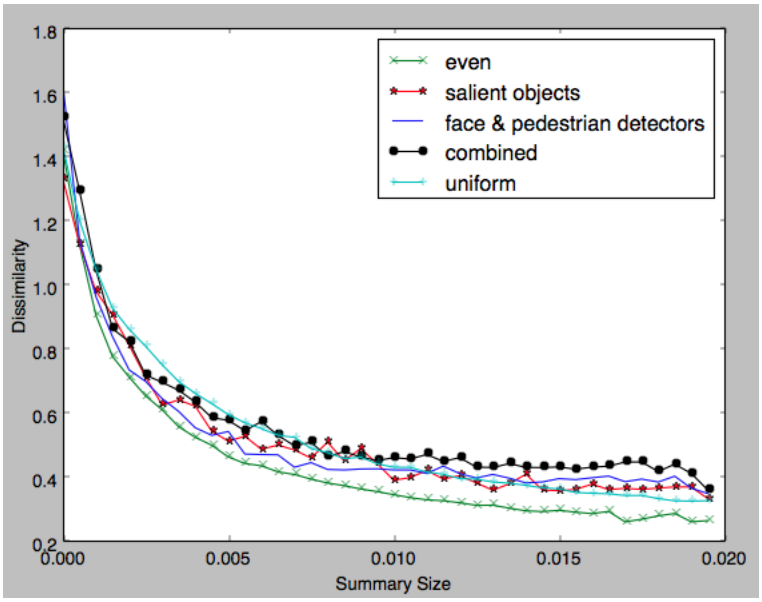

Figure 6.15: Dissimilarity Curves (Important-People-Centered)

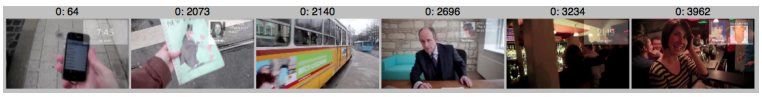

Figure 6.16: 6-Frame Keyframe Summary: "Uniform" Method

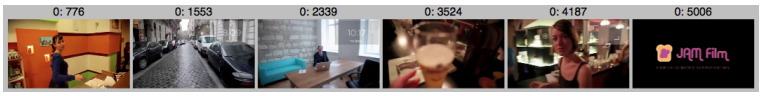

Figure 6.17: 6-Frame Keyframe Summary: "Evenly-Spaced" Method

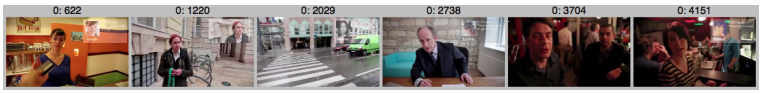

Figure 6.18: 6-frame Keyframe Summary: "Face and Pedestrian Detectors" Method

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 7:
Discussion

This chapter discusses how the results presented in the previous chapter support the achievement of the author's thesis goals, which are (1) the development of video summarization methods that surpass baseline methods and (2) the development of an evaluation framework that can provide quantitative measures of keyframe summary quality that existing evaluation frameworks fail to provide.

7.1 Thesis Goal 1

The keyframe summary evaluation results for each summarization ideal (scenery-type-centered, face-centered, and important-people-centered) are discussed in the context of thesis goal (1) below. In general, the author shows that better summaries can be constructed with high-level, albeit imperfect, content detectors than by employing baseline summarization methods.

For the discussion, the author assumes that the ROC metric is the primary measure of keyframe summary quality. In terms of the stakeholder's summarization ideal, it reveals the extent to which the desired content is present in a keyframe summary, for a given amount of redundancy. The other metrics (variability and dissimilarity) reveal how the desired content is distributed within a keyframe summary. So, the author assumes that capturing the desired content is more important than capturing the desired content in the "right" proportions. However, the evaluation framework presented allows one to weight the proposed metrics in any desirable fashion.

7.1.1 Scenery Types

The "Scenery Transition Peaks" algorithm achieved the best ROC score. This means that it produced keyframe summaries with the most desirable content, for a given amount of content redundancy. More specifically, the "Scenery Transition Peaks" algorithm produced keyframe summaries that capture content between indoor, outdoor, natural, man-made, and greenery scenery video transitions better than baseline methods. These results demonstrate

that the object bank scene classification technology developed by Li et al. [43] may be employed in a keyframe summarization context to achieve better-than-baseline outcomes.

Qualitatively observing the keyframe summary produced by the "Uniform" method in Figure 6.4, it appears that more redundancy and less desirable content occurs than does in the keyframe summary produced by the "Scenery Transition Peaks" method in Figure 6.6. The results are less striking when comparing the "Scenery Transition Peaks" keyframe summary in Figure 6.6 with the "Evenly-Spaced" keyframe summary in Figure 6.5. The narrow difference between the "Scenery Transition Peaks" ROC curve and the "Evenly-Spaced" ROC curve in Figure 6.1 may account for the less dramatic qualitative difference. Future work may employ a more diverse set of scenery classifiers to improve upon the results.

In terms of the distribution of the desirable scenery content, the "Evenly-Spaced" baseline algorithm produced keyframe summaries that are both more variable and less dissimilar. Once again, these are considered secondary measures of keyframe summary quality, so the results for these measures do not affect the overall algorithm assessment.

7.1.2 Faces

The "Face Frames Color Histogram Clustering" algorithm achieved the best ROC score. This means that it produced keyframe summaries better than baseline methods, in that, they possess more content from the different face portions of a given video. Of note, the "Face Segments" and "Face Peaks" methods also outperform the baseline methods. The results demonstrate that face detection technology may be employed in a keyframe summarization context to achieve better-than-baseline outcomes.

Qualitatively observing the keyframe summaries produced by the "Uniform" method in Figure 6.10, the "Evenly-Spaced" method in Figure 6.11, and the "Face Frames Color Histogram Clustering" method in Figure 6.12, the "Face Frames Color Histogram Clustering" keyframe summary is clearly superior. The wide margin between the "Face Frames Color Histogram Clustering" ROC curve and the baseline ROC curves in Figure 6.7 may explain the dramatic qualitative difference.

Viola-Jones face detection technology was employed in the "Face Frames Color Histogram Clustering," "Face Segments," and "Face Peaks" algorithms. Future work may employ

more sophisticated face-tracking technologies, such as that developed by Sivic et al. [40] to improve upon the results.

7.1.3 Important People

The "Face and Pedestrian Detectors" algorithm achieved the best ROC score. This means that it produced keyframe summaries better than baseline methods, in that, they possess more important people, in their different environments, in different size groups. The "Combined" approach also outperformed the baseline methods. The results demonstrate that face, pedestrian, and salient object detection technologies may be employed in a keyframe summarization context to achieve better-than-baseline outcomes.

Qualitatively observing the keyframe summaries produced by the "Uniform" method in Figure 6.16, the "Evenly-Spaced" method in Figure 6.17, and the "Face and Pedestrian Detectors" method in Figure 6.18, the "Face and Pedestrian Detectors" keyframe summary possess the most important people. Although the visual results presented here may not provide definitive evidence of the superiority of the "Face and Pedestrian Detectors" algorithm, especially when compared to the "Evenly-Spaced" keyframe summary, the ROC curves presented in Figure 6.13 do.

Lee et al. [34] also demonstrate keyframe summarization technology that outperforms baseline methods. In particular, the authors attempt to generate good keyframe summaries, with respect to a video's important objects. To this end, they develop "important-object" detection technology. They compare their keyframe summarization method to a uniform sampling method and show that it is superior, that is, for any summary size, it finds a larger percentage of the important objects. Future work may expand the AOC from important people to important objects and compare the methods developed in this paper with those developed by Lee et al.

7.2 Thesis Goal 2

The keyframe summary evaluation results are discussed in the context of thesis goal (2) below. The author shows that the evaluation framework developed in this work provides quantitative measures of keyframe summary quality that existing evaluation frameworks fail to provide.

The scenery-centered results show the extent to which a keyframe summary captures video content between indoor, outdoor, natural, man-made, and greenery scenery transitions. The face-centered results show the extent to which a keyframe summary captures content from the temporally different face portions of a video. These measurements are unique mainly because they are absent from previous works. Scenery classification technologies exist, but they are neither employed nor evaluated in the context of keyframe summarization [42,43]. Sivic et al. [40] developed sophisticated face tracking technology, but, once again, they neither employ nor evaluate it in the context of keyframe summarization.

The measurements presented in this paper are also unique because they are semantically high-level. In other words, they reveal keyframe summary quality with respect to semantically high-level stakeholder concerns, such as scenery and objects (faces, important people). Many other measurements of keyframe summary quality are semantically low-level, that is, keyframe summary quality is defined as a function of low-level visual characteristics of the keyframe set. Semi-Hausdorff distances, shot reconstruction degree (SRD), distortion, and energy scores have been utilized in this regard [23, 24, 46, 47]. These semantically low-level measurements of keyframe summary quality may be of limited use to a stakeholder who is interested in high-level concepts, such as objects and events.

Keyframe summary quality measures exist that are semantically high-level. For example, Lee et al. [34] present such a quantitative measure with respect to important objects. Here, a good summary is one that captures a high percentage of a video's important objects. Our work also presents keyframe summary quality measures that are semantically high-level. For example, the important-people-centered results presented in our paper show the extent to which a keyframe summary captures all of a video's important persons, in their different environments, in different size groups. This measure is novel because, when compared to existing measures, it reveals keyframe summary quality with respect more complex stakeholder summarization concerns.

Consider an important object, a dog, for example. Consider a video that has a number of different breeds of dogs. One stakeholder may desire content from each portion of the video where any dog is simply present. Another stakeholder's concerns may be more refined. They may desire content from each portion of the video where a particular breed of dog is present. Our evaluation framework captures this concern with properties of concern

(POCs). For example, the important-people-centered results reveal keyframe summary quality with respect to different important persons in different environments. Here, the POCs are a person's identity and their environment.

Additionally, consider an important object, a gun, for example. Consider a video that has a number of distinct segments where the same particular gun is present and absent. One stakeholder may value content from each segment where the particular gun is present. Another stakeholder may desire content from only a single segment where the gun is present. Our evaluation framework captures these stakeholder concerns with temporal concerns (TCs). For example, the face-centered results reveal the extent to which a keyframe summary captures content from the temporally different face portions of a video.

Finally, a stakeholder may consider portions of a video with many important objects to be distinct from portions of a video with few important objects. In this case, content from video portions with different numbers of important objects would be desirable in the keyframe summary. Our evaluation framework captures this stakeholder concern with combinatorial concerns (CCs). The important-people-centered results reveal keyframe summary quality with respect to different group sizes, for example.

Existing high-level evaluation frameworks do not incorporate the above concerns (TCs, POCs, and CCs) into their ground truth summaries, so their measures of keyframe summary quality do not reveal the extent to which these concerns are captured [34, 35, 48].

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 8:
Conclusions

The evaluation framework presented in this work connects a stakeholder's semantically high-level keyframe summarization needs to quantitative measures of performance. This connection was missing from previous keyframe evaluation methods and is crucial to development and assessment of keyframe summarization systems.

Computer vision technologies continue to improve and can automatically recognize semantically high-level visual concepts, such as objects and events. Some of these technologies are ripe for application to the problem of automatic keyframe video summarization. However, blind application of these powerful technologies is problematic. The quality of a keyframe summary is not an intrinsic property, but rather, it exists in the context of a stakeholder's needs. The evaluation framework presented in this work enables measuring the quality of a keyframe summary with respect to a stakeholder's semantically high-level summarization needs. Consequently, stakeholders and algorithm developers can now identify when they have achieved success. This is also crucial to the technological progress of the field. Truong and Venkatesh [5] identified the need for large comparison studies in the field of keyframe summarization. Without these studies it is difficult to evaluate and select the most promising automatic keyframe summarization technologies and methods. The evaluation framework presented in this work enables such comparison studies. This work showed that, with an agreed upon summarization ideal, numerous and varied keyframe summarization technologies can be meaningfully compared to each other based upon their quantitative metric achievements.

In addition to allowing comparison of one algorithm to another, the metrics presented in this work can be used by stakeholders to set software performance thresholds, that is, the metrics can be used to set product acceptance criteria.

Finally, this work further confirmed that, when it comes to semantically high-level summarization concerns, content-dependent methods can outperform baseline or content-independent approaches [34].

THIS PAGE INTENTIONALLY LEFT BLANK

REFERENCES

[1] D. Vesset, B. Woo, H. D. Morris, R. L. Villars, G. Little, J. S. Bozman, L. Borovick, C. W. Olofson, S. Feldman, and S. Conway, "Worldwide big data technology and services 2012–2015 forecast," International Data Corporation, Framingham, MA, Tech. Rep. 233485, 2012.

[2] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers, "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, New York, NY, Tech. Rep., 2011.

[3] Youtube. (2014). Statistics. [Online]. Available: http://www.youtube.com/yt/press/statistics.html.

[4] "Cisco visual networking index: Forecast and methodology, 2012–2017," Cisco Systems, San Jose, CA, Tech. Rep. 11684, May 2013.

[5] B. T. Truong and S. Venkatesh, "Video abstraction: A systematic review and classification," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 3, no. 1, Feb. 2007.

[6] B. C. O'Connor, "Selecting key frames of moving image documents: A digital environment for analysis and navigation," *Microcomputers for Information Management*, vol. 8, no. 2, pp. 119–133, 1991.

[7] E.-K. Kang, S. J. Kim, and J.-S. Choi, "Video retrieval based on key frame extraction in compressed domain," in *Proceedings. 1999 International Conference on Image Processing*, vol. 3, 1999, pp. 260–264 vol.3.

[8] H. J. Zhang, J. Wu, D. Zhong, and S. W. Smoliar, "An integrated system for content-based video retrieval and browsing," *Pattern Recognition*, vol. 30, no. 4, pp. 643 – 658, 1997.

[9] M. Yeung and B. Liu, "Efficient matching and clustering of video shots," in *Proceedings. 1995 International Conference on Image Processing*, vol. 1, Oct 1995, pp. 338–341 vol.1.

[10] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proceedings. 1998 International Conference on Image Processing*, vol. 1, Oct 1998, pp. 866–870 vol.1.

[11] W. Xiong, C.-M. Lee, and R.-H. Ma, "Automatic video data structuring through shot partitioning and key-frame computing," *Machine Vision and Applications*, vol. 10, no. 2, pp. 51–65, 1997.

[12] R. Hammoud and R. Mohr, "A probabilistic framework of selecting effective key frames for video browsing and indexing," in *International Workshop on Real-Time Image Sequence Analysis (RISA'00)*, 2000, pp. 79–88.

[13] A. Ferman and A. Tekalp, "Two-stage hierarchical video summary extraction to match low-level user browsing preferences," *Multimedia, IEEE Transactions on*, vol. 5, no. 2, pp. 244–256, June 2003.

[14] A. Divakaran, K. Peker, R. Radhakrishnan, Z. Xiong, and R. Cabasson, "Video summarization using mpeg-7 motion activity and audio descriptors," in *Video Mining*, ser. *The Springer International Series in Video Computing*, A. Rosenfeld, D. Doermann, and D. DeMenthon, Eds. Springer US, 2003, vol. 6, pp. 91–121.

[15] X.-D. Yu, L. Wang, Q. Tian, and P. Xue, "Multilevel video representation with application to keyframe extraction," in *Proceedings. 10th International Conference on Multimedia Modelling*, Jan 2004, pp. 117–123.

[16] L. He, E. Sanocki, A. Gupta, and J. Grudin, "Auto-summarization of audio-video presentations," in *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, ser. MULTIMEDIA '99. New York, NY, USA: ACM, 1999, pp. 489–498.

[17] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in *Proceedings of the Tenth ACM International Conference on Multimedia*, ser. MULTIMEDIA '02. New York, NY, USA: ACM, 2002, pp. 533–542.

[18] N. BABAGUCHI, Y. KAWAI, and T. KITAHASHI, "Generation of personalized abstract of sports video," *2012 IEEE International Conference on Multimedia and Expo*, vol. 0, p. 158, 2001.

[19] H. Sundaram and S.-F. Chang, "Condensing computable scenes using visual complexity and film syntax analysis," in *PROCEEDINGS OF ICME 2001*, 2001, pp. 389–392.

[20] J. quan Ouyang, J.-T. Li, and Y.-D. Zhang, "Replay boundary detection in mpeg compressed video," in *2003 International Conference on Machine Learning and Cybernetics*, vol. 5, Nov 2003, pp. 2800–2804 Vol.5.

[21] A. Rav-Acha, Y. Pritch, and S. Peleg, "Making a long video short: Dynamic video synopsis," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2006, pp. 435–441.

[22] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[23] T. Liu, X. Zhang, J. Feng, and K.-T. Lo, "Shot reconstruction degree: a novel criterion for key frame selection," *Pattern Recognition Letters*, vol. 25, no. 12, pp. 1451 – 1457, 2004.

[24] H.-C. Lee and S.-D. Kim, "Iterative key frame selection in the rate-constraint environment," *Signal Processing: Image Communication*, vol. 18, no. 1, pp. 1 – 15, 2003.

[25] V. Kobla, D. DeMenthon, and D. S. Doermann, "Special-effect edit detection using videotrails: a comparison with existing techniques," vol. 3656, 1998, pp. 302–313.

[26] J. Yuan, H. Wang, L. Xiao, W. Zheng, J. Li, F. Lin, and B. Zhang, "A formal study of shot boundary detection," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 2, pp. 168–186, Feb 2007.

[27] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," vol. 41, no. 6, pp. 797–819, 2011.

[28] C. Liu, "Beyond pixels : exploring new representations and applications for motion analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 2009.

[29] R. Laganière, R. Bacco, A. Hocevar, P. Lambert, G. Païs, and B. E. Ionescu, "Video summarization from spatio-temporal features," in *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, ser. TVS '08. New York: ACM, 2008, pp. 144–148.

[30] A. Girgensohn and J. Boreczky, "Time-constrained keyframe selection technique," in *1999 IEEE International Conference on Multimedia Computing and Systems*, vol. 1, Jul 1999, pp. 756–761 vol.1.

[31] X.-D. Yu, L. Wang, Q. Tian, and P. Xue, "Multilevel video representation with application to keyframe extraction," in *Proceedings. 10th International Conference on Multimedia Modelling*, Jan 2004, pp. 117–123.

[32] D. Gibson, N. Campbell, and B. Thomas, "Visual abstraction of wildlife footage using gaussian mixture models and the minimum description length criterion," in *Proceedings. 16th International Conference on Pattern Recognition*, vol. 2, 2002, pp. 814–817 vol.2.

[33] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra, "Adaptive key frame extraction using unsupervised clustering," in *Proceedings. 1998 International Conference on Image Processing*, vol. 1, Oct 1998, pp. 866–870.

[34] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1346–1353.

[35] F. Dufaux, "Key frame selection to represent a video," in *Proceedings. 2000 International Conference on Image Processing*, vol. 2, Sept 2000, pp. 275–278 vol.2.

[36] H.-W. Kang and X.-S. Hua, "To learn representativeness of video frames," in *Proceedings of the 13th Annual ACM International Conference on Multimedia*, ser. MULTIMEDIA '05. New York, NY, USA: ACM, 2005, pp. 423–426.

[37] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient Object Detection: A Discriminative Regional Feature Integration Approach," *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2083–2090, Jun. 2013.

[38] P. Wang, J. Wang, G. Zeng, J. Feng, H. Zha, and S. Li, "Salient object detection for searched web images via global saliency," in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3194–3201.

[39] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," *Comput. Vis. Image Underst.*, vol. 81, no. 3, pp. 231–268, Mar. 2001.

[40] J. Sivic, M. Everingham, and A. Zisserman, "Person spotting: Video shot retrieval for face sets," in *Proceedings of the 4th International Conference on Image and Video Retrieval*, ser. CIVR'05. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 226–236.

[41] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[42] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 15, no. 1, pp. 52–64, Jan 2005.

[43] L.-J. Li, H. Su, E. P. Xing, and L. Fei-fei, "Object bank: A high-level image representation for scene classification and semantic feature sparsification," 2010.

[44] L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *2007 IEEE 11th International Conference on Computer Vision*, Oct 2007, pp. 1–8.

[45] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 413–420.

[46] H. S. Chang, S. Sull, and S.-U. Lee, "Efficient video indexing scheme for content-based retrieval," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 9, no. 8, pp. 1269–1279, Dec 1999.

[47] T. Liu and J. Kender, "Optimization algorithms for the selection of key frame sequences of variable length," in *Computer Vision — ECCV 2002*, ser. Lecture Notes in Computer Science, A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, Eds. Springer Berlin Heidelberg, 2002, vol. 2353, pp. 403–417.

[48] C. Kim and J.-N. Hwang, "Object-based video abstraction using cluster analysis," in *Proceedings. 2001 International Conference on Image Processing*, vol. 2, Oct 2001, pp. 657–660 vol.2.

[49] Merriam-Webster. (2014). artifact. [Online]. Available: http://www.merriam-webster.com/dictionary/artifact.

[50] L. Freeman, *Elementary applied statistics: for students in behavioral science*. Wiley, 1965.

[51] A. R. Wilcox, "Indices of qualitative variation and political measurement," *The Western Political Quarterly*, vol. 26, no. 2, pp. pp. 325–343, 1973.

[52] C. E. Shannon, "Prediction and entropy of printed english," *Bell System Technical Journal*, vol. 30, pp. 50–64, Jan. 1951.

[53] Youtube. (2014). [Online]. Available: http://www.youtube.com/.

[54] C. Vondrick, D. Ramanan, and D. Patterson, "Efficiently scaling up video annotation with crowdsourced marketplaces," in *Computer Vision – ECCV 2010*, ser. Lecture Notes in Computer Science, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Springer Berlin Heidelberg, 2010, vol. 6314, pp. 610–623.

[55] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, May 2001.

[56] Imagenet. (2014). [Online]. Available: image-net.org.

[57] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[58] M. Kolsch. (2014). Easy! Computer Vision. [Online]. Available: https://github.com/NPSVisionLab/CVAC.

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California